Changing for the Better

Preference Dynamics and Agent Diversity

Fenrong Liu

Changing for the Better

Preference Dynamics and Agent Diversity

ILLC Dissertation Series DS-2008-02

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
e-mail: illc@science.uva.nl
homepage: http://www.illc.uva.nl/

Changing for the Better

Preference Dynamics and Agent Diversity

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Aula der Universiteit
op dinsdag 26 februari 2008, te 12.00 uur

door

Fenrong Liu

geboren te Shanxi, China.

Promotor: Prof.dr. J.F.A.K. van Benthem
Co-promotor: Prof.dr. D.H.J. de Jongh

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

This research was financially supported by the Universiteit van Amsterdam and the Netherlands Organization for Scientific Research (NWO).

Cover design by Meng Li, Tsinghua University, Beijing.
Printed and bound by PrintPartners Ipskamp

ISBN: 978–90–5776–175–1

谨以此书献给我亲爱的母亲罗猫英和我亲爱的父亲刘永才!

Contents

Acknowledgments

It has been quite a journey to finish this thesis, with so many people to thank.

I thank my promotor, Johan van Benthem, who brought me back to Amsterdam when I thought I would never have the chance. He has walked with me as I struggled to create something new over all these years. His vision, guidance, and encouragement helped me in all the research and writing that went into this thesis. I have always been overwhelmed with his incredibly stimulating comments, and every visit to his office reminded me of what I once read in Nottingham University: "Arrive alert, leave enlightened". I learned a lot, not only in logic, but also in many other aspects. My appreciation for his trust and constant support is beyond any words that I could possibly find. He is one of the rare supervisors that students dream to have. I was lucky indeed.

I thank my co-promotor, Dick de Jongh, for all the attention that he has paid to me during these years. I feel grateful for his help in the process of my Mosaic grant application, which indirectly led to my current project, funded by the Dutch Science Organization NWO and the University of Amsterdam. I am constantly impressed by his instantaneous grasp of difficult problems. He is always ready to help. I learned much from him, especially through working on our joint paper. His careful reading of all of my writing and his corrections will never be forgotten. In particular, I deeply appreciate his never-failing understanding and kindness. I also thank him for teaching me his vast knowledge on biking.

I thank Natasha Alechina, Alexandru Baltag, Vincent Hendricks, John-Jules Meyer, and Frank Veltman for their willingness to assess my dissertation.

I thank the following people with whom I have discussed questions concerning my thesis in person or via email. Their views have shaped my thoughts: Thomas Ågotnes, Natasha Alechina, Guillaume Aucher, Reinhard Blutner, Jan Broersen, Denis Bonnay, Hans van Ditmarsch, Paul Egré, Ulle Endriss, Patrick Girard, Till Grüne-Yanoff, Sven Ove Hansson, Paul Harrenstein, Andreas Herzig, Brian Hill, Alistair Isaac, Jérôme Lang, Brian Logan, Ondrej Majer, Teresita Mijangos, Eric Pacuit, Rohit Parikh, Leo Perry, Adrian Piper, Floris Roelofsen, Josh Snyder,

Leon van der Torre, Tomoyuki Yamada, and Henk Zeevat.

ILLC is an ideal place to study. Yde Venema, Maricarmen Martínez, Frank Veltman, Krister Segerberg, Johan van Benthem, and Peter van Emde Boas taught me a lot in the courses I took from them. In addition, many results of this thesis were first presented at the 'dynamics seminar', whose participants gave me very helpful feedback. I want to thank: Harald Bastiaanse, Tijmen Daniels, Cédric Dégremont, Ulle Endriss, Patrick Girard, Sujata Ghosh, Fernando Velazquez-Quesada, Sieuwert van Otterloo, Eric Pacuit, Floris Roelofsen, Olivier Roy, Yanjing Wang, and Jonathan Zvesper.

ILLC is also a unique place as a friendly cultural environment. I thank my officemate Brian Semmes for sharing the office, and for his effort to speak "zhoumo yukuai". I thank the following persons who taught me about different cultures: Nick Benzhanishvili, Gaëlle Fontaine, Amélie Gheerbrant, Nina Gierasimczuk, Tikitu de Jager, Joost Joosten, Aline Honingh, Daisuke Ikegami, Clemens Kupke, Lena Kurzen, Olivia Ladinig, Markos Mylonakis, Alessandra Palmigiano, Yoav Seginer, Merlijn Sevenster, Leigh Smith, Reut Tsarfaty, Sara Uckelman, Joel Uckelman, Raul Leal Rodriguez, Jakub Szymanik, Levan Uridia, Jacob Vosmaer, Andreas Witzel, and Jelle Zuidema.

Balder ten Cate, Loredana Afanasiev, Maarten Marx, Yoav Seginer, and Sonja Smets provided essential encouragement during the hard time of my final thesis writing. Patrick Girard read my Chapter 4 and provided very useful comments. Natasha Alechina, Dick de Jongh, Stefan Minica, and Yanjing Wang helped improve the typographical presentation. Olivier Roy helped arrange the defense and many other logistic things.

Special thanks go to Ingrid van Loon, Tanja Kassenaar, Marjan Veldhuisen, Jessica Pogorzelski, Karin Gigengack, Peter van Ormondt, René Goedman, and Adri Bon for their being so nice and warm. They were always immediately ready to help with any administrative or other non-scientific problem.

My Chinese friends in Amsterdam have been a big family for me. Whenever I need a hand, they are always available. I thank them for their help and entertainment: Bo Gao, Hai Hu, Yuli Huo, Fangbin Liu, Huiye Ma, Zhenwei Pu, Bosheng Tsai, Xiaoping Wei, Yanjing Wang, Yingying Wang, Hang Wang, Mengxiao Wu, Ming Wu, Jie Xie, Fan Yang, Junwei Yu, Shibiao Zhang. In particular, I want to thank Zhisheng Huang and his family for all their care these years.

I also want to thank my colleagues in Beijing: Junren Wan, Xiaochao Wang, and Lu Wang at Tsinghua University, Jingyuan Li, Shangmin Wu, Dikun Xie, and Qingyu Zhang at the Chinese Academy of Social Sciences, for their generous consideration in allowing me to finish my thesis abroad.

Finally, I thank my husband Xiaowei Song for his quiet support and patience that enabled me to complete this work.

Fenrong Liu
Amsterdam, January, 2008.

Chapter 1

Introduction

Preference is what colors our view of the world, and it drives the actions that we take in it. Moreover, we influence each other's preferences all the time by making evaluative statements, uttering requests, commands, and statements of fact that exclude or open up the possibility of certain actions. A phenomenon of this wide importance has naturally been studied in many disciplines, especially in philosophy and the social sciences. This dissertation takes a formal point of view, being devoted to logical systems that describe preferences, changes in preference, and behaviors of different agents in dynamic contexts. I will plunge right in, and immediately draw your attention to the first time when preference was fully discussed by a logician.

Preference logic in the literature

What von Wright considered, and what he did not In his seminal book *The Logic of Preference: An Essay* from 1963, von Wright started with a major division among the concepts that interest moral philosophers. He divided them into the following three categories (though there may be border-line cases):

- *deontological* or *normative*: notions of right and duty, command, permission and prohibition,
- *axiological*: notions of good and evil, the comparative notion of betterness,
- *anthropological*: notions of need and want, decision and choice, motive, end and action.

The intuitive concept of preference itself was said to 'stand between the two groups of concepts': It is related to the axiological notion of betterness on one side, but it is related just as well to the anthropological notion of choice.

While considering the relationship between preference and betterness, von Wright distinguished two kinds of preference relations: *extrinsic* and *intrinsic* ones. He explains the difference with the following example:

> "... a person says, for example, that he prefers claret to hock, because his doctor has told him or he has found from experience that the first wine is better for his stomach or health in general. In this case a *judgement of betterness serves as a ground or reason* for a preference. I shall call preferences, which hold this relationship to betterness, *extrinsic*.
>
> It could, however, also be the case that a person prefers claret to hock, not because he thinks (opines) that the first wine is better for him, but simply because he likes the first better (more). Then his liking the one wine better is not a reason for his preference. ..."
>
> ([Wri63], p.14)

Simply stated, the difference is principally that p is preferred *extrinsically* to q if it is preferred *because* it is better in some explicit respect. If there is no such reason, the preference is intrinsic.

Instead of making the notion of betterness the starting-point of his inquiry,[1] von Wright took a more "primitive" intrinsic notion of preference as 'the point of departure', providing a formal system for it which has generated a whole subsequent literature (cf. [Han01a]).

We are by no means claiming that the division between intrinsic and extrinsic preference is the only natural way of distinguishing preferences. One can also study varieties of moral preference, aesthetic preference, economic preference, etc. However, in this thesis, I will follow von Wright's distinction. Our first main goal is to extend the literature on intrinsic preferences with formal logical systems for the *extrinsic notion of preference*, allowing us to spell out the reasons for a preference. On the way there, we will also make new contributions to the literature on intrinsic preferences.

Besides the extrinsic notion of preference that was removed from von Wright's agenda, there is another important issue which he left open. More precisely, he writes the following:

> "The preferences which we shall study are a subject's intrinsic preferences on one occasion only. Thus we exclude both *reasons* for preferences and the possibility of *changes* in preferences."
>
> ([Wri63], p.23)

Clearly, our preferences are not static! One may *revise* one's preferences for many legitimate (and non-legitimate) reasons. The second main issue dealt with in this thesis is how to model preference change in formal logics. This leads to new dynamic versions of existing preference logics, and interesting connections with belief revision theory.

[1] [Hal57] did propose logic systems for the notion of betterness.

What others considered afterwards, and what they did not Following von Wright's work, many studies on preference were carried out over the last few decades. Due to its central character, at the interface between evaluation, choice, action, moral reasoning, and games, preference has become a core research theme in many fields, which have often led to logical theory. In what follows I will summarize the main issues or directions taken by other researchers. My purpose is not to give an overview of the vast literature (I give some basic references for that), but only to point out some issues that are relevant to the present thesis, and some particular proposals that have inspired it.

Preference in logic and philosophy Formal investigations on preference logic have been mainly carried out by philosophical logicians. The best survey up to 2001 can be found in the Chapter *Preference Logic* by Sven Ove Hansson in the *Handbook of Philosophical Logic*.

This literature added several important notions to von Wright's original setting. In particular, a distinction which has played an important role is that between preference over *incompatible* alternatives and preference over *compatible* alternatives, based on early discussions in [Wri72]. The former is over *mutually exclusive* alternatives, while the latter does not obey this restriction. Here is a typical example:

> "In a discussion on musical pieces, someone may express preferences for orchestral music over chamber music, and also for Baroque over Romantic music. We may then ask her how she rates Baroque chamber music versus orchestral music from the Romantic period. Assuming that these comparisons are all covered by one and the same preference relation, some of the relata of this preference relation are not mutually exclusive."
>
> ([Han01a], p.346-347)

Most philosophical logicians have concentrated on exclusionary preferences. However, in this thesis we will consider both. As we will see, one of our logical systems is for preference over objects, which are naturally considered as exclusive incompatible alternatives. But we will also work with preferences between propositions, which can be compatible, and indeed stand in many diverse relationships.

Also, most researchers have been particularly interested in the question whether certain *principles* or 'structural properties' are reasonable for preference. Here economists joined logicians, to discuss the axioms of rational preference. Many interesting examples have been proposed to argue for or against certain formal principles, resulting in different logical systems (cf. [Tve69], [Sch75], [Lee84], etc.). However, a general critical result in [Han68] is worth being noticed. In this paper, the author showed that many axioms proposed for a general theory of preference imply theorems which are too strange to be acceptable. But it is often possible to

restrict their domain of application to make them more plausible. In general, our logical systems will not take a strong stand on structural properties of preference, beyond the bare minimum of reflexivity and transitivity (though we note that the latter has been questioned, too: Cf. [Hug80], [Fis99]).

There are also obvious relationships between preference and *moral* or more generally, *evaluative* notions like "good" and "bad". Several researchers have suggested definitions for "good" and "bad" in terms of the dyadic predicate "better". A widespread idea is to define "good" as "better than its negation" and "bad" as "worse that its negation", as in [Wri63] and [Hal57].[2] Alternatively, [CS66b] presents indifference-related definitions for "good" and "bad", and then defines things as follows: "a state of affairs is good provided it is better than some state of affairs that is indifferent, and ... a state of affairs is bad provided some state of affairs that is indifferent is better than it". [Han90a] generalized the previous proposals, and presented a set of logical properties for "good" and "bad". Interestingly, precisely the opposite view has been defended in the logical literature on semantics of natural language. [Ben82] defines binary comparatives like "better" in terms of context-dependent predicates "good", and [Roo07] takes this much further into a general analysis of comparative relations as based on a 'satisfying' view of achieving outcomes of actions.[3] Either way, we will not pursue this particular line of analysis in this thesis, although one might say that our later analysis of preference as based on constraints has some echoes of the linguistic strategy deriving binary comparatives from unary properties.

The connection between preference and moral reasoning is clear in *deontic logic*, another branch of philosophical logic going back to von Wright's work, this time to [Wri51]. While obligation is usually explained as truth in all 'deontically accessible worlds', the latter are really the 'best worlds' in some moral comparison relation. Not surprisingly, then, preference relations were introduced in standard deontic logic to interpret conditional obligations. For modern preference-based deontic logics, see [Han90b], [Tor97]. Preference was introduced particularly to help solve some of the persistent 'deontic paradoxes'. Here are a few examples: [CS66a] gave a moral deontic interpretation of the calculus of intrinsic preference, to solve the *problem of supererogation* - 'acting beyond the call of duty'.[4] [TT98] extended the existing temporal analysis of Chisholm's Paradox of conditional obligation (see [Eck82], too) using a deontic logic that combines temporal and preferential notions. Also, [TT99] provided better solutions to many paradoxes by combining preferential notions with *dynamic updates*: making this dynamics even more explicit will be one of our main themes.

[2] Quantitative versions of these ideas are found in [Len83].

[3] It would be of interest to contrast their formal 'context-crossing principles' with Hansson's proposals.

[4] Non-obligatory well-doing is traditionally called supererogation. Many of the great deeds of saints and heroes are supererogatory.

Preference in decision theory and game theory The notion of preference is also central to decision theory and game theory: given a set of feasible actions, a rational agent or player compares their outcomes, and takes the action that leads to the outcome which she most prefers. Typically, to make this work, outcomes are labeled by quantitative utility functions - though there are also foundational studies based on qualitative preference ordering ([Han68]). Moving back to logic, [Res66] brought together the concepts of preference, *utility* and of *cost* that play a key role in the theoretical foundations of economics, studying primarily the metric aspect of these concepts, and the possibility of measuring them. For modern discussions in this line, see [Bol83] and [Tra85]. In terms of axiomatization, the standard approach takes weak preference ("better or equal in value to") as a primitive relation, witness [Han68] and [Sen71].

In particular, economists have studied connections between *preference* and *choice* ([Sen71], [Sen73]), treating *preference* as almost identical with *choice*. Preference is considered to be 'hypothetical choice', and choice to be *revealed preference*. Recently, revealed preference has become prominent in understanding the concept of equilibrium in game theory (cf. [HK02]). Differently from standard logical models, preference is then attached to observed outcomes. Preferences of players have to be constructed, so that the observed outcomes can be rationalized by the chosen equilibrium notion employing these constructed preferences.

But, one has to be careful with such identifications of notions across different fields. Preference is not really the same as choice. Many researchers have remarked on that. Already in [Wri63], it was pointed out that 'it is obvious that there can exist intrinsic preferences, even when there is no question of *actually* choosing between things."([Wri63], p.15). Choice must involve actual action, but preference need not. In this thesis, we will not pursue the connection between preference and its emergence in general action, though our dynamic framework for describing preference change can presumably be extended to deal with the latter scenario.[5]

Preference in computer science and AI From the 1980s onward, and especially through the 1990s, researchers in computer science and AI have started paying attention to preference as well. Their motivations are clear: 'agents' are central to modern notions of computation, and agents reason frequently about their preferences, desires, and goals. Thus, representing preferences and goals for decision-theoretic planning has become of central significance. For instance, [CL90] studied general principles that govern agents' reasoning in terms of their belief, goals and actions and intentions. The well-known 'BDI model' was first presented in [RG91] to show how different types of rational agents can be modeled by imposing conditions on the persistence of an agent's beliefs, desires or

[5] A related area of formal studies into preferences for agents, and how these can be merged, is *Social Choice Theory*: Cf. [Fis73].

intentions, and its further development can be found in [LHM96], [HW03], and [Woo00]. Other work in qualitative decision theory illustrates how planning agents are driven by goals (defined as desires together with commitments) performing sequences of actions to achieve these (cf. [Bou94], [DT99], [Tho00]). Of interest to logicians, general properties of the language of preference representation have become important, such as striking a balance between expressive power and succinctness (see [CMLLM04] and [CEL06]).[6]

Further occurrences of preference are found in the AI literature on common sense reasoning, witness the treatment of circumscription, time, 'inertia', and causality in [Sho88]. Interestingly, further crucial notions from von Wright have made their way directly into this literature. In particular, his idea that preferences can often only be stated *ceteris paribus* has been taken up in [DSW91] and [DW94], which studied preference "all else being equal". The other main sense of 'ceteris paribus', as "all else being normal", was taken up in [Bou93], where preference relations are based on what happens in the most likely or "normal" worlds. A recent development of ceteris paribus preference in a modal logic framework is [BRG07]. The eventual systems of our thesis can deal with the latter, though not (yet) with the former.

Some specific influences on this dissertation

In terms of new methods for the logic of preferences, we now mention a few sources here that have influenced this dissertation. The authors in [LTW03] propose a logic of desires whose semantics contains two ordering relations of preference and normality, respectively. They then interpret desires as follows: "in context A, I desire B" iff "the best among the most normal $A \wedge B$ worlds are preferred to the most normal $A \wedge \neg B$ worlds". Such combinations are typical of what we will deal with eventually. But before getting to these entangled scenarios, we also employ tools from straight preference logic, in particular, the *modal preference logics* proposed by [Bou94], and following him, [Hal97]. Halpern started with just a betterness ordering over possible worlds, and showed how to extend this to sets of possible worlds. He then gave a complete axiomatization of this logic over partial orders. This sets the model for the basic 'static' completeness results we will need later.

But there are yet more influences on our work from the computational literature. One obvious one is *propositional dynamic logic* for sequential programs and general actions ([HKT00]), which will be our main model for describing the dynamics of preference change. Our semantics and complete logics will follow especially the modern format of *dynamic epistemic logic* ([DHK07]). We will

[6] Indeed, preferences are also found in the more 'hard core' theory of computation, e.g., in describing evolutions of computational systems, which need to be compared as to some measure of 'goodness'. Substantial examples of this trend are [Mey96] on dynamic logic with preference between state transitions, and [Ser04] on a general calculus of system evolution.

elaborate on this paradigm in more detail below, and in the main body of the thesis. But there are even further sources. Interestingly, the recent computational literature also takes up themes from social choice theory, such as *aggregation of preferences*, as a matter of crucial interest to describing the behavior of societies of agents, cooperative or competing. One sophisticated study of this sort, which brings together social choice theory, preference logic, and algebraic logic, is [ARS02]. We will use their techniques for preference merge triggered by hierarchies of agents to shed light on the array of notions involved in intrinsic and extrinsic preferences.

This concludes our survey of major developments in preference logics as relevant to this thesis. The account is by no means complete, however. For instance, many further connections between preference, belief revision, and the foundations of *economics* are found in [Rot01]. And also, it will soon be clear that our treatment of extrinsic preferences, generated by further outside considerations, also owes much to *linguistics*, viz. the area of Optimality Theory ([PS93]), which describes grammatical sentences and successful utterances in a rule-free manner, in terms of optimal satisfaction of syntactic, semantic, and pragmatic *constraints*. How constraints induce preference, and how they can enter preference logic, will be a major theme in what follows.[7] But for now, we summarize where we stand.

Our starting point is the preference logic of von Wright, and some major distinctions that he made. We identified two major issues that [Wri63] left out, viz. *reason-based extrinsic preference*, and the *dynamics of preference change*. Of course, we are not claiming that nobody paid any attention to these two issues over the past decades. But it does seem fair to say that most authors took the notion of intrinsic preference only, and concentrated on its properties.[8] Next, we have only found a few papers treating changes in preference as such. [BEF93] is a first attempt at using dynamic logic for this purpose. Also, influenced by *AGM*-style belief revision theory, [Han95] proposed postulates for four basic operations in preference change.

Against this background, this thesis will show how these two crucial aspects of reasoning with preference can be treated in a uniform logical framework, which borrows ideas from several different areas: (a) the subsequent development of preference logic, (b) the computational literature on agents, (c) linguistic optimality theory, and (d) recent developments in the theory of belief revision and dynamic epistemic logic.

Having reviewed what has been done by others, here is what is new in this thesis. Basically, I will study a number of old issues that are still open, and a

[7] These ideas are even extended into models for brain function in cognitive science (cf. [Smo04]).

[8] Still, 'reasons for preference' are a theme in decision theory and economics, witness the brief survey in [HGY06].

few new issues that have not yet been considered. Also, I will study these issues only from a *formal logical point of view*. In what follows I introduce my guiding intuitions, and the main ideas.

On intuitions and ideas

Reasons for preference In many situations, it is quite natural to ask for a reason when someone states her preference to you. It may be a matter of justification for her, but as for you, you simply want more explanation or information (sometimes, in order to judge whether it is rational for her to have that preference). So preference can come with a reason, and this is what von Wright called 'extrinsic preference'. Let us return to the example used by [Wri63] to explain this notion:

> A person prefers claret to hock, *because* his doctor has told him or he has found from experience that the first wine is better for his stomach or health in general.

Here, *the first wine being better for his health* is the reason for his preference of claret to hock. Similar examples abound in real life: one prefers some house over another *because* the first is cheaper and of better quality than the second.

Conceptually, reasons stand at a different level from preferences, and they form a *base* or *ground* for their justification. Reasons can be of various kinds: from general principles to more 'object-oriented' facts. In many cases, one can combine more than one reason to justify one single preference. Thus, in the house example, not only the price of the house matters, but also the quality. In such cases, reasons may have their own structure, and different considerations may be ordered according to their importance. One may think for instance that the quality of a house is more important than its price.

Preference change There is more to be said about the above example. Let us first add a twist of imagination to make it dynamical:

> Suppose that before he sees the doctor, he *preferred hock to claret*. Now the doctor tells him "the first wine is better for your health". He then *changes* his preference, and will now *prefer claret to hock*!

Again such things often occur in real life. We change our preferences on the basis of new information that we have received. And the new preference emerges for a new reason. Actually, this way of thinking immediately links us to information dynamics in general. Accordingly, I will use the methodology of modeling information dynamics to deal with preference change in this dissertation. A few more words are in order here. The idea behind information dynamics is this: agents receive new information and update their knowledge or beliefs accordingly. This

style of thinking can be traced back to the early 1980s, e.g., the well-known AGM postulates handling belief change ([AGM85]). But the approach I am taking here is what recently developed under the name of *dynamic epistemic logic* (DEL). It has a certain Amsterdam flavor, which inspired me through the following works: [Pla89], [Vel96], [Ben96], [BMS98], [Ger99], and [DHK07], as well as up-to-date work on belief revision by [Ben07a] and [BS08]. Readers will see in the later chapters how I apply techniques from these works to the dynamics of preference. This choice of approach also distinguishes my proposals from the AGM-style preference change presented in [Han95].

We know that reasons and preferences live at different levels. Moreover, reasons provide an explanation for preference. Thus one can travel between the two levels, as reasons lead to a preference, and preference can be seen as derived from reasons. Since dynamics can take place at both levels, we will also investigate how to relate the changes at the two levels to each other.

Beliefs as a reason, too There is one issue we have not yet considered in the above, namely, *uncertainties*. When someone tries to give a reason for her preference, in some situations, she may not have precise information to offer. Instead, she may say things like 'I believe that it is going to rain, so I prefer bringing my umbrella'. Under such circumstances, one's preference relies on one's *beliefs*, and beliefs come in as an extra reason for preference. People may have different preferences *because* they have different beliefs. Thus the notion of preference becomes richer, and similarly changes in preference as well: preference change may now also be caused by belief change.

The literature on preference logic has not considered intertwined belief and preference yet. But such entanglements are standard in other areas, in particular, decision theory, which has a tradition in modeling decision making under uncertainty ([Sav54], [Jef65]). Here most models rely on a numerical representation where utility and uncertainty are commensurate. For instance, an agent may not know the outcomes of his actions, but may use a probability distribution over outcomes instead. The expected value of an action can be then computed from utility and probability, as explained in any textbook. What is relevant to our preceding discussion is this. The main reason to represent worlds probabilistically in decision theory is to be able to use the *beliefs* as a base for decision making. By contrast, we will use beliefs as well, but mostly in a qualitative approach without numerical calculations.[9]

We are diverse human beings Preferences notoriously differ, and this variety seems typical of human behavior and interaction. But this diversity extends to other features of agent behavior. For instance, consider the reasons people have for preferences, and the ways these might change. Here, too, different people

[9] We refer to [Liu06b] for some numerical counterparts to our qualitative proposals.

may react quite differently. In particular, when belief is involved, this naturally leads to various policies for changing beliefs, a diversity which is at the heart of belief revision theory. For instance, a 'radical agent' may change her preference immediately, taking her reasons from some partial information received, whereas a 'conservative agent' will stick to past beliefs and past preferences for longer, waiting for more input. In addition to these differences in preference, and belief policies, agents may also have differences in their even more basic logical capacities for information handling: in particular, their memory capacity, and tendencies to forget crucial information obtained earlier.

This diversity of agent behavior seems an essential fact of life to us, and one of the most striking features human interaction is how it still leads to coordinated, and often very successful behavior. Such phenomena have been studied to some extent in belief revision theory, witness the host of belief revision policies in [Rot06]. Another area of diversity studies is in models for inferential information and computational restrictions on agents abilities (cf. [BM07], [Egr04], [BE07]). But these aspects of diversity have not yet been studied in their totality, and we will make an attempt in this thesis to provide a more comprehensive model of agents whose preferences, beliefs, and information may vary, as well as the dynamic rules which change these.

Given these considerations, the challenge is to understand how, despite our differences, we live in one society, interacting with each other successfully. The following questions then arise:

- What major aspects can agents differ in?
- How differently do they update their knowledge and beliefs when facing new information?
- How do they interact with each other, say in games, despite these differences - say, by learning each other's 'type' of behaviour?

These questions have come up in several areas. For instance, game theorists have studied 'bounded rationality' in the study of cooperative behavior([OR94], [Axe84]), while, as we said, formal epistemologists have tried to parameterize agents' inferential or computational powers. Moreover, the variety of human behavior versus idealized norms has been emphasized in the study of reasoning in cognitive psychology ([Gol05], [HHB07]).

But more in particular, the preceding questions pose a serious challenge to the dynamic logics for preference change and belief that we have developed. Do they leave room for significant differences in agent behavior across the appropriate range of variation that can be found in practice? It may seem that they do not, since 'the valid reduction axioms' for knowledge after update seem written in stone. Even so, this thesis will show that dynamic logics do allow for the proper variation, by providing formal logical models for variety inside dynamic epistemic

logic, which address the preceding issues as well as others. We will relate them to the study of games, and general processes in a temporal setting.

Connections to related areas

As stated at the outset, this thesis is squarely within the logical tradition. Nevertheless, beyond obvious comparisons to be made with the older literature on preference logic, I believe that my results may be of interest to some of the other areas mentioned here. For instance, the qualitative perspective on preference and preference change may be of interest to decision theorists looking for qualitative models. Likewise, since preference comes with its own intuitions, dynamic logics of preference can be inspirational for dynamic logics of beliefs. A case in point is [BS08] whose account of new belief modalities and reduction axioms for them was influenced by [BL07]. Furthermore, since the models proposed here are abstract and general, they can be applied to neighbors of preference logic such as deontic logic. I believe that norm change and obligation change can be modeled in a similar way to preference change, and indeed, a number of such studies have been made, including [Zar03], and in particular, a recent series of papers by Tomoyuki Yamada, of which [Yam07] is a representative sample. Indeed, vice versa, their work has also influenced mine. Finally, in the philosophy of action, our treatment of the difference between intrinsic preference and extrinsic preference may provide a synthesis between so-called "recognitional" and "constructivist" views of practical reasoning.[10] Our two notions of preference explain such a difference in a precise way. For further connections between preference logic and the philosophy of action, see the two dissertations [Gir08], [Roy08] which touch this one at various points mentioned in subsequent chapters.

Finally, I will briefly state the structure of the thesis in slightly more technical terms, showing how my intuitions and ideas are formalized in logics.

Structure of the thesis

This thesis is organized as follows:

In Chapter 2, a first model for extrinsic preferences is proposed. Models consist of a universe of possible worlds, representing the different relevant situations, endowed with a basic objective order of 'betterness'. The latter supplies 'reasons for preference' when we 'lift' this order to one among propositions, viewed as sets of possible worlds.[11] There are various kinds of 'lifting', of which we consider in

[10] According to the "recognitional" view, rational practical reasoning consists in trying to figure out which of the available options are good things to do, and then choosing accordingly. According to the "constructivist" view, rational practical reasoning consists in complying with certain conditions of purely formal coherence or procedural rationality. For more details on the debate, see [Wed03].

[11] Note that betterness is a preference over incompatible alternatives.

particular the $\forall\exists$-version saying that every φ world has at least one better ψ alternative world. These lifts, and many other types of statement can be described in a standard modal language over betterness models. As for the dynamics of preference change, this is triggered as follows. Statements like suggestions or commands 'upgrade' agents' current preferences by changing the current betterness order among worlds. A complete logic of knowledge update plus preference upgrade is presented that works with dynamic-epistemic-style reduction axioms. The result is an intertwined account of changing preferences and also changing knowledge as triggered by factual information. This system can also model changing obligations, conflicting commands, or 'regret' about possibilities that have already been ruled out epistemically. Beyond specific examples, we present a general format of relation transformers for which dynamic-epistemic reduction axioms can be derived automatically.

Chapter 3 provides a second model for extrinsic preferences. This time, the aim is to analyze preferences over objects, again, comparing incompatible alternatives. For this purpose, inspired by linguistic optimality theory, the primary structure is an ordered 'priority sequence' of 'constraints', i.e., relevant properties of objects. It supplies reasons for preference by comparing objects as to the properties they have or lack in this sequence. Typically, the relationship between reasons and the resulting extrinsic preference is characterized by so called 'representation theorems' in this chapter.[12] Intuitively, these results say that one can always find a reason for some given object preference. Next, in the realistic case where agents only have incomplete information, here, too, we add epistemic structure. In particular, we introduce beliefs that help form preferences. Three definitions are proposed to describe how different kinds of agents get their preference under uncertainties. Changes of preference are then explored with two different reasons: either changes in the priority sequence, and also through belief change. Both can lead to preference change.

In Chapter 4, I primarily draw a comparison between the two approaches in Chapters 2 and 3, both qua semantics and qua syntax. First, abstract *structured models* are introduced to merge 'reasons' (a set of ordered propositions) and a correlated 'betterness order' over possible worlds. I then study general ways of deriving world preferences from an ordered set of propositions, as well as the opposite direction: ways of lifting a world preference relation to an ordering over propositions. Interestingly, when we go back and forth between these, we find several tight correspondences between concrete order-changing operations at the two levels, and some specific definability results are proved. The general context behind these are partially ordered 'priority graphs' from the literature on preference merge, which seem the most elegant mathematical framework behind our specific proposals. We prove definability results at this level, too, and draw a comparison with 'priority product update' in the dynamic epistemic logic of

[12] As usual, these results may be viewed as structural versions of completeness theorems.

belief revision. Then, we briefly look at the different formal languages used in our various systems, and contrast and compare them. Next, in line with both Chapters 2 and 3, I extend the setting from pure preference to intertwining of preference, knowledge, and beliefs. Several new concepts of preference will be defined, in a sequence of modal languages of ascending strength. Finally, I compare with a new proposal of combining all systems studied so far into one grand 'doxastic preferential predicate logic' of both object and world preference.

In Chapter 5, I move to a setting where habits of preference and belief change are just one aspect of general diversity of agents. Agents are *not* all the same, and nevertheless, they manage to coordinate with each other successfully. I start with the observation that dynamic epistemic logic presupposes that every agent remembers all the actions she has taken before. But then I show that this is a negotiable assumption, which can be dropped from the framework. In particular, *memory-bounded agents* are defined and their behavior is captured in a new dynamic epistemic completeness theorem with a key reduction axiom different from the usual one. Next, following ideas from my master of logic thesis [Liu04], I consider different policies in belief revision, and suggest how a continuum of these, too, can be incorporated into dynamic epistemic logic. These logics allow for co-existence of different memory capacities and revision policies, and hence, through different modal operators, they can describe the interplay of diverse agents. Throughout the chapter, imperfect information games, viewed as finite trees of possible actions with epistemic uncertainties, are used as a playground.

Finally, in Chapter 6, diversity of agents is discussed in a more abstract and systematic way. The major sources of diversity are considered first, such as inferential powers, introspective ability, powers of observation, memory capacity, and revision policies. I then show how these can be encoded in dynamic epistemic logics allowing for individual variation among agents along many dimensions. Furthermore, I explore the interaction of diverse agents by looking at some concrete scenarios of communication and learning. A logical methodology to deal with these issues is proposed as well.

Chapter 7 concludes the dissertation and identifies some major further issues for research that come to light once we put our chapters together into one account of diverse preference-driven agents.

Origins of the material Material from these chapters has been presented at several colloquia and conferences, including ESSLLI 2005 (Edinburgh), ESSLLI 2006 (Malaga), LOFT 2006 (Liverpool), and Luxembourg Workshop on Norm Change 2007. As for publications, Chapter 2 is the published joint paper ([BL07]). Chapter 3 is an extension of the joint paper ([JL06]) as submitted for publication organized after the *Workshop on Modeling Preference Change* in Berlin, 2006. Chapter 4 is largely new, and partly a product of the 'dynamics seminar' at ILLC Amsterdam. Chapter 5 is an updated and extended version of the published joint

paper ([BL04]). Chapter 6 is an extension of the accepted paper ([Liu06a]) of the *Workshop on Logics for Resource Bounded Agents* in Malaga, 2006, and it will appear in the *Journal of Logic, Language and Information*.

Chapter 2

Dynamic Logic of Preference Upgrade

2.1 Introduction: changing preferences

The notion of preference occurs across many research areas, such as philosophy of action, decision theory, optimality theory, and game theory. Individual preferences between worlds or actions can be used to predict behavior by rational agents. More abstract notions of preference also occur in conditional logic, non-monotonic logic and belief revision theory, whose semantics order worlds by relative similarity or plausibility.

Preference logics Preference logics in the literature describe different comparative structures by means of various devices ([Han90b]). Agents' preferences can run between worlds or between actions, preference statements can be weaker or stronger in what they say about worlds or actions being compared – and also, they may be more 'objective' or more 'epistemic'. A statement like "I prefer sunsets to sunrises" can be cast merely in terms of 'what is better for me', or as a more complex propositional attitude involving my beliefs about the relevant events. In this chapter, we take an objective approach, where a binary preference relation supports a unary modality "true in some world which is at least as good as the current one" ([Bou94], [Hal97]). [BOR06] show how such a language, when extended with a few operators from hybrid languages, can define several conditionals, Nash equilibrium, and backward induction solutions to games. The language also expresses various kinds of preference that agents may have between propositions, i.e., types of events. Moreover, we add explicit epistemic operators, allowing us to express agents' attitudes toward what is good or better for them.

Preference dynamics Our main concern in this chapter, however, is one of *dynamics*. Preferences are not static, but they change through commands of moral authorities, suggestions from friends who give good advice, or just changes

in our own evaluation of worlds and actions. Such changes can have various triggers. For instance, intuitively, a command

"See to it that φ!"

makes worlds where φ holds preferred over those where it does not - at least, if we accept the preference induced by the issuer of the command. But also a process of planning, with just our own goals in mind, may gradually introduce preferences over actions as ways toward reaching the goal, as we learn more about the actual world. These and other dynamic aspects of preference have been noted by many authors, including [BEF93], [Han95], [Zar03], [TT99], and [Yam06].

Related ideas all play in the dynamic semantics for conditional logics (for instance, [Spo88], [Vel96]). In its static Lewis-style semantics, a conditional $\varphi \Rightarrow \psi$ says roughly the following

ψ is true in all most-preferred φ-worlds $\quad (\natural)$

But one plausible way of accepting a conditional is, not as a true/false description of a current preference, but rather as an instruction for *adjusting* that preference so as to make $(\natural)$ the case. Even more simply, consider a default assertion like

"*Normally* φ."

As [Vel96] points out, this does not eliminate $\neg\varphi$-worlds from our current model, in the usual dynamic sense of information update. Accommodating this assertion rather makes the $\neg\varphi$-worlds doxastically less preferred than φ-worlds.

Trigger 1: suggestions There are many triggers for preference change in real life, and dynamic preference logics should provide a format for studying these in an appropriate generality. To find such formats, in this chapter, we start from a simple test scenario that may be called a 'suggestion'. Consider someone who is indifferent between taking a trip (p) and staying at home ($\neg p$). Now his friend comes along and says

"Let's take a trip!"

'Taking' this suggestion means that any preference we might have had for staying at home is removed from the current model. Figure 2.1 shows what we have in mind.

Thus, in our scenario, a suggestion removes already existing preference links: but it does not add new ones. Note that, in addition to arrows drawn, our preference relations always have reflexive loops. This mechanism will be studied in greater detail later on, as an entry into more general kinds of preference upgrade. Even so, by way of contrast, here is one alternative, which does not remove links, but rather adds them.

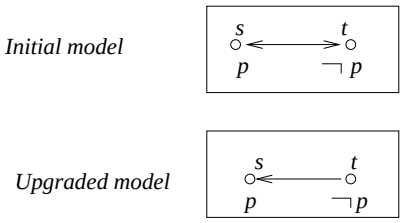

Figure 2.1: Initial model and upgraded model

Trigger 2: commands In the above picture, the agent now prefers the trip, so this has become her priority, or in a deontic reading of the preference relation, her duty. But in general, suggestions are weaker than commands. Taking the suggestion does not mean that the person will now prefer all p-worlds to the $\neg p$-ones. It all depends on the preference structure already in place. If the agent was indifferent between p and $\neg p$ with arrows both ways, the suggestion induces a preference. But the agent may be unable to compare the two situations, as in this model with two unrelated worlds:

Figure 2.2: A model with two unrelated worlds

A suggestion in the relation-decreasing sense does not make the worlds comparable. With real commands "Take that trip!", however, we want to make sure the agent now prefers p. Then, we need to *add* preference links to the picture, making the world with $\neg p$ less preferred. Our proposals also deal with upgrades that add links between worlds.

Dynamic logics of upgrade Whether eliminative or additive, preference change is reminiscent of existing systems for information *update* in dynamic-epistemic logic ([Ger99], [BMS98], [Ben07a], [DHK07]). In the latter paradigm, incoming assertions or observations change the domain of the current model and/or its accessibility relations. In our scenario, current preference relations are changed by incoming suggestions or commands. Thus, we will speak henceforth of preference *upgrade* as a counterpart to the better-known term update. The main point of this chapter is that preference upgrade is a viable phenomenon, just as susceptible to systematic modification as information, temporal perspective, or other parameters of 'logical dynamics' ([BEF93], [Ben96], or in the setting of conditional logic, [Spo88], [Vel96]). We will show how this dynamics can be implemented by the very same methodology that has been developed for information update in dynamic-epistemic logic.

This chapter is structured as follows. First we present a new joint epistemic preference logic (Section 2.2). Its semantics is based on preferences between worlds. This allows us to talk about knowing or not knowing one's preferences, or regretting that the best scenario is not going to happen. Next, in Section 2.3, we provide formal definitions for preference upgrade, with an emphasis on the above 'suggestions' increasing our preference for one proposition over its negation. Interestingly, this also suggests alternative formulations for information update. Section 2.4 defines a dynamic version of the static epistemic preference language, where information update lives together with preference upgrade. There is a completeness theorem in terms of the usual style of reduction axioms recursively analyzing postconditions of actions. This is our first 'existence proof' for a compositional dynamics of upgrade, in tandem with update of information. In Section 2.5, we consider more general upgrade scenarios: first with general schemes of link elimination, and then, with the full strength of 'product update' for information using 'event (action) models'. This requires enriching the action models of dynamic-epistemic logic with agents' preferences between events. Section 2.6 then outlines some applications of our dynamic upgrade logics, to default reasoning, deontic logic, and logics of commands. Section 2.7 is a brief survey of related work, and Section 2.8 contains our conclusions and further directions.

This chapter proposes a certain style of thinking about preference upgrade, and an existence proof for a logical methodology in doing so. We do not address all intuitive senses of preference, or all logical issues arising in the areas where it plays a role. A more extensive discussion of upgrade mechanisms with various triggers, various senses of preference, and further applications, is found in [JL06] and later chapters in this thesis.

2.2 Epistemic preference logic

Language and semantics

The main language used in this chapter has two components: a preference modality as in [BOR06], and the standard knowledge operators from epistemic logic.

2.2.1. Definition. Take a set of propositional variables P and a set of agents I, with p ranging over P and i over I. The *epistemic preference language* is given by the following rule:

$$\varphi ::= \bot \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_i\varphi \mid [pref]_i\varphi \mid U\varphi.$$

Intuitively, $K_i\varphi$ stands for 'agent i knows that φ', while $[pref]_i\varphi$ says that all worlds which the agent considers as least as good as the current one satisfy φ. U is an auxiliary universal modality.[1]

[1]For technical convenience, we often shift to the corresponding existential modalities $\langle K\rangle_i$, $\langle pref\rangle_i$, and $E\varphi$. These seem more difficult to read in terms of intuitive linguistic expressions.

How is this formal language connected to 'preference' as it occurs in natural discourse? One may be inclined to read $\langle pref\rangle_i\varphi$ as 'agent i prefers φ'. But as with other logical systems, there is a gap between the formalism and common usage. E.g., just saying that the agent sees some better world where φ holds seems too weak, while the universal modality $[pref]_i\varphi$ 'in all better worlds' seems much too strong. Cf. [Han01b] for a thorough discussion of senses of preference, and ways in which formal languages do or do not match up. Here we just point out the following facts. First, our formal language can also express intermediate senses of 'betterness' for preference, using *combinations* of modalities. E.g., $[pref]_i\langle pref\rangle_i\varphi$ will express, at least on finite connected models, that some *best* world has φ. And one can also express that *all* best worlds satisfy φ: cf. [BOR06]. Moreover, our approach emphasizes comparisons of worlds, i.e., objects, rather than propositions, whereas common notions of preference often play between propositions, or semantically, sets of worlds. Such preferences between propositions can be *defined* on our approach (see again [BOR06]). For instance,

$$U(\psi \rightarrow \langle pref\rangle_i\varphi)$$

expresses one strong sense of 'agent i prefers φ to ψ', viz. each ψ-world s has at least one epistemic alternative which is φ and which is at least as good as s according to the agent. But one can also define the original notion of preference in [Wri63] which says that the agent prefers all φ-worlds to all ψ-worlds (cf. [BOR06]; [BRG07] also deals with Von Wright's 'ceteris paribus' clause in the relevant comparisons between worlds). For the moment, we take this expressive power of our simple-looking modal language for granted. The virtue of our simple base modalities is that these 'decompose' more complex preference statements in a perspicuous manner, while allowing for a simple dynamic approach later on.

2.2.2. Definition. An *epistemic preference model* is a tuple $\mathcal{M} = (S, \{\sim_i|\ i \in I\}, \{\preceq_i|\ i \in I\}, V)$, with S a set of possible worlds, $\sim_i$ the usual equivalence relation of epistemic accessibility for agent i,[2] and V a valuation for proposition letters. Moreover, $\preceq_i$ is a reflexive and transitive relation over the worlds.

We read $s \preceq_i t$ as 't is at least as good for agent i as s', or 't is weakly preferred to s'. If $s \preceq_i t$ but not $t \preceq_i s$, then t is *strictly preferred* to s, written as $s \prec_i t$. If $s \preceq_i t$ and $t \preceq_i s$, then agent i is *indifferent* between s and t. Models can also have a distinguished actual world, but we rarely use this feature here.

Note that we do not require that our preference relations be *connected* in the sense of the Lewis sphere models for conditional logic. In general, we want to

But they help in finding and checking valid principles, and in semantic arguments generally.

[2]Interpreting the knowledge operator with the equivalence relation is optional in an approach. There are many philosophical discussions about its justification. Various alternatives haven been proposed in terms of model classes. For complete epistemic logics over equivalence relations or other model classes, see the standard references, e.g. [FHMV95] or [BRV01].

allow for genuinely incomparable worlds where an agent has no preference either way, not because she is indifferent, but because she has no means of comparing the worlds at all. This is just as in the semantics for the minimal conditional logic. Of course, in special settings, such as the standard utility-based preference orderings of outcomes in a game, connectedness may be quite appropriate.

2.2.3. Definition. Given an epistemic preference model $\mathcal{M} = (S, \{\sim_i | \ i \in I\}, \{\preceq_i | \ i \in I\}, V)$, and a world $s \in S$, we define $\mathcal{M}, s \models \varphi$ (formula φ *is true in* $\mathcal{M}$ *at* s) by induction on φ:

1. $\mathcal{M}, s \models p$ iff $s \in V(p)$.
2. $\mathcal{M}, s \models \neg\varphi$ iff not $\mathcal{M}, s \models \varphi$.
3. $\mathcal{M}, s \models \varphi \wedge \psi$ iff $\mathcal{M}, s \models \varphi$ and $\mathcal{M}, s \models \psi$.
4. $\mathcal{M}, s \models \langle K \rangle_i \varphi$ iff for some $t : s \sim_i t$ and $\mathcal{M}, t \models \varphi$.
5. $\mathcal{M}, s \models \langle pref \rangle_i \varphi$ iff for some $t : s \preceq_i t$ and $\mathcal{M}, t \models \varphi$.
6. $\mathcal{M}, s \models E\varphi$ iff for some t: $\mathcal{M}, t \models \varphi$.

Expressive power As we noted, [BOR06] have shown that the pure modal preference part of this language, with the help of the universal modality, can express a variety of natural notions of preference between propositions, including the original one proposed by Von Wright, as well as other natural options qua quantifier combinations. Moreover, following [Bou94], they show that this language can faithfully embed non-iterated conditionals $\varphi \Rightarrow \psi$ using the above preference operator $\langle pref \rangle_i$, as follows:

$$U(\varphi \rightarrow \langle pref \rangle_i (\varphi \wedge [pref]_i (\varphi \rightarrow \psi))).$$

But with our additional epistemic operators, we can also express the interplay of preference and knowledge. The following examples represent (a) an intuition of self-reflection of 'preference', and (b) an unfortunate but ubiquitous phenomenon:

- $\langle pref \rangle_i \varphi \rightarrow K_i \langle pref \rangle_i \varphi$: Preference Positive Introspection
- $\langle pref \rangle_i \varphi \wedge K_i \neg\varphi$: Regret.

We will return to mixed epistemic-preference principles later on.

Proof system and completeness

Our epistemic preference logic can be axiomatized completely in a standard modal style, given our choice of epistemic preference models (cf. [BRV01]).

2.2.4. THEOREM. *Epistemic preference logic is completely axiomatizable w.r.t epistemic-preference-models.*

Proof. The proof is entirely by standard techniques. □

Additional axioms in our language impose further frame conditions on models. Here are two examples to show the spirit. They are based on standard modal frame-correspondence techniques:

2.2.5. FACT.

- A preference frame $\mathcal{F} = (S, \{\sim_i | \ i \in I\}, \{\preceq_i | \ i \in I\})$ satisfies *connectedness*, i.e., $\forall x \forall y : x \preceq_i y \vee y \preceq_i x$, iff the following formula is true in the frame:

  $$(\varphi \wedge E\psi) \rightarrow \langle pref \rangle_i \psi \vee E(\psi \wedge \langle pref \rangle_i \varphi).$$

- An epistemic preference frame $\mathcal{F}$ makes the *Preference Introspection Axiom* $\langle pref \rangle_i \varphi \rightarrow K_i \langle pref \rangle_i \varphi$ true iff it satisfies the following condition:

  $$\forall s \forall t \forall u : (s \preceq_i t \wedge s \sim_i u \rightarrow u \preceq_i t).$$

Nevertheless, we will work with the minimal system described above in this chapter, leaving such extras to asides.

2.3 Modelling preference upgrade

Brief review of epistemic information update

The basic paradigm for epistemic update is public announcement. Suppose that an agent does not know if p is the case, but learns this fact through an announcement $!p$. Then we get the following sort of model change pictured in Figure 2.3, where the dotted line in the initial static model indicates the agent's uncertainty in the initial situation.

The announcement eliminates the $\neg p$-world from the epistemic model, and afterwards, the agent knows that p. There is an extensive literature on dynamic epistemic logics for public announcements and more sophisticated epistemic events, that can modify information in different ways for different agents. See [BMS98], [Ben07a], and Section 2.5 below.

These logics all work essentially on the same design principle. First, a class of models is chosen representing the relevant information structures, together with

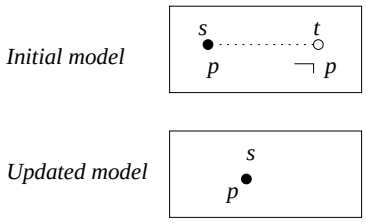

Figure 2.3: Initial model and updated model

some appropriate static language for describing these. Usually, these are models for some version of standard epistemic logic. Next, an update mechanism is proposed which transforms given models under some chosen set of epistemic actions. For public announcement, this simply means a possible world elimination, yielding a definable submodel:

> *A public announcement* $!\varphi$ *of a true proposition* φ *turns the current model* $(\mathcal{M}, s)$ *with actual world* s *into the model* $(\mathcal{M}_{!\varphi}, s)$ *whose worlds are just the set* $\{w \in S \mid \mathcal{M}, w \models \varphi\}$*. And accessibility relations and valuations are retained on the restricted domain.*

More complex actions update to *products* $\mathcal{M} \times \mathcal{E}$ of the current epistemic model $\mathcal{M}$ with some 'event model' $\mathcal{E}$ containing all relevant events or actions.

Next, the static language gets a dynamic extension where the informative events themselves are displayed and manipulated. For public announcement, a typical static-dynamic assertion of this sort is

$[!\varphi]K_i\psi$*: after a truthful public announcement of* φ*, the agent* i *knows that* ψ*.*

Here the semantic clause for the dynamic modality is simply as follows:

$$\mathcal{M}, s \models [!\varphi]\psi \text{ iff } (\text{if } \mathcal{M}, s \models \varphi, \text{ then } \mathcal{M}_{!\varphi}, s \models \psi).$$

Usually, the effects of events can then be described completely in a recursive manner, leading to a compositional analysis of communication and other cognitive processes. As a crucial illustration, here is the key *reduction axiom* in current logics of public announcement for a true assertion resulting in an epistemic possibility for agent i:

$$\langle !\varphi\rangle\langle K\rangle_i\psi \leftrightarrow \varphi \wedge \langle K\rangle_i\langle !\varphi\rangle\psi.$$

As discussed in the literature, semantically, this reflects a sort of perfect recall for updating agents. Computationally, axioms like this help drive a reduction algorithm for dynamic epistemic statements to static epistemic statements, allowing us to borrow known decision procedures for the base language.

Upgrade as relation change

With the paradigm of public announcement in mind, we now define the mechanism of preference change described informally in the above. Our static models are of course the epistemic preference structures of Section 2.2:

$$\mathcal{M} = (S, \{\sim_i |\ i \in I\}, \{\preceq_i |\ i \in I\}, V)$$

Our triggers are events of publicly suggesting φ, written as follows:

$$\sharp\varphi$$

These lead to the following model change, removing preferences for $\neg\varphi$ over φ:

2.3.1. Definition. Given any epistemic preference model $(\mathcal{M}, s)$, the *upgraded model* $(\mathcal{M}_{\sharp\varphi}, s)$ is defined as follows.

(a) $(\mathcal{M}_{\sharp\varphi}, s)$ has the same domain, valuation, epistemic relations, and actual world as $(\mathcal{M}, s)$, but

(b) the new preference relations are now

$\preceq_i^* = \preceq_i - \{(s,t) \mid \mathcal{M}, s \models \varphi \text{ and } \mathcal{M}, t \models \neg\varphi\}$.[3]

We suppress agent subscripts henceforth whenever convenient.

Upgrade for suggestion events replaces a preference relation by a definable subrelation. This may be written as follows in the standard notation of propositional dynamic logic (e.g. [HKT00]):

$$R := R - (?\varphi; R; ?\neg\varphi).$$

We will consider more general relation-changing operations in Section 2.5. For instance, if one wanted to add links, rather than just subtract them, the format would still work. E.g., the relation-extending stipulation

$$R := R \cup (?\neg\varphi; \top; ?\varphi),$$

where $\top$ is the universal relation, would make every φ-world preferable to every $\neg\varphi$-world. With our upgrade defined, we are in a position to define a dynamic language for preference upgrade. But before doing so in Section 2.4, we consider some features of the mechanism just defined.

[3] [Har04] analyzes newly defined preference relations in a set-theoretic format.

Preservation properties of upgrade Perhaps the most pressing issue is whether a proposed model changing operation stays inside the class of intended static models. For the update associated with public announcements $!\varphi$, this was so - and the reason is the general logical fact that submodels preserve *universally defined* relational properties like reflexivity, transitivity, and symmetry. For our notion of upgrade, the properties to be preserved are reflexivity and transitivity of preference relations (epistemic relations remain unchanged). This time, no general result comes to the rescue, since we only have the following counterpart to the preservation result for submodels:

2.3.2. FACT. The first-order properties preserved under taking subrelations are precisely those definable using *negated atoms*, $\wedge$, $\vee$, $\exists$, $\forall$.

But neither reflexivity nor transitivity is of this particular syntactic form. Nevertheless, using some special properties of our proposal, we can prove

2.3.3. FACT. The operation $\mathcal{M}_{\sharp\varphi}$ preserves reflexivity and transitivity.

Proof. Reflexivity is preserved since we never delete loops (s, s). As for transitivity, suppose that $s \preceq^* t \preceq^* u$, while not $s \preceq^* u$. By the definition of $\sharp\varphi$, we must then have $\mathcal{M}, s \models \varphi$ and $\mathcal{M}, u \models \neg\varphi$. Consider the intermediate point t. Case 1: $\mathcal{M}, t \models \varphi$. Then the link (t, u) should have been removed from $\preceq$. Case 2: $\mathcal{M}, t \models \neg\varphi$. In this case, the link (s, t) should have been removed. Either way, we have a contradiction. □

On the other hand, our upgrades $\sharp\varphi$ can lead to loss of connectedness of the preference order. Our earlier example already showed this in Section 2.1 (see Figure 2.2). Likewise, our upgrades can lead to a loss of positive introspection, see the following scenario:

2.3.4. EXAMPLE. Consider Figure 2.4 below. There are two worlds 'asleep' and 'awake'. In both models, we do not know if we are sleeping or awake. Initially, we prefer being asleep, and we know our preference. Now an upgrade happens, suggesting that real waking life is not so bad after all. Then we still do not know if we are sleeping or awake, but at the 'awake' world we prefer being awake (thought not to be the case at the 'asleep' world). Focusing on the 'asleep' world in the new model, we still prefer being asleep there. But we no longer know that we prefer it – since we might be in the 'awake world'. Introspection fails!

In some settings, preference introspection seems plausible, and a desirable property of models to be preserved. We can then change the above notion of upgrade to deal with this, e.g., by making sure that similar links are removed at epistemically indistinguishable worlds, or study which special sorts of upgrade in our language have the property of always preserving preference introspection. The latter would then be the 'reasonable' or 'sensible' series of suggestions.

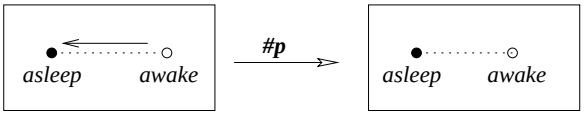

Figure 2.4: Upgrade leading to a loss if positive introspection

Update by link cutting Update and upgrade do not lead wholly separate lives in our setting. For instance, if we want to model the earlier phenomenon of 'regret' about worlds that are no longer viable options, epistemic updates for $!\varphi$ should not remove the $\neg\varphi$-worlds, since we might still want to refer to them, and perhaps even mourn their absence. One way of doing this is by redefining the update for public announcement as a relation-changing operation of 'link cutting'. This time, instead of the above $!\varphi$, we write the relevant update action as follows, note that the notation '!' is now behind φ:

$$\varphi!$$

and we write the updated model as $\mathcal{M}_{\varphi!}$ in order to distinguish it from what we obtain by eliminating worlds. We should really change notations to reflect the two kinds of exclamation mark – but we trust the reader can disambiguate in context. The correct semantic operation for $\varphi!$ on models is this:

2.3.5. Definition. The *modified public update model* $\mathcal{M}_{\varphi!}$ is the original model $\mathcal{M}$ with its worlds and valuation unchanged, but with accessibility relations $\sim_i$ replaced by a version without any crossing between the φ- and $\neg\varphi$-zones of $\mathcal{M}$:

$$(?\varphi; \sim_i; ?\varphi) \cup (?\neg\varphi; \sim_i; ?\neg\varphi)$$

2.3.6. Fact. The pure epistemic logic of public announcement is the same with $!\varphi$ and with $\varphi!$.

Nevertheless, the second update stipulation has some advantages. It was first proposed, in [Sny04] (cf. Chapter 5 and 6) for modelling the behavior of *memory-free* agents, whose epistemic accessibility relations are quite different from those for the idealized update agents of standard dynamic epistemic logic. Moreover, in the present setting, in stating regrets, we need the consistency of a formula like

$$K_i p \wedge \langle pref \rangle_i \neg p.$$

Yes, I know that p, but it would be better if it weren't... Modified update allows us to have this consistently.

Link cutting has some curious features, too. E.g., link cutting in the current model is the same for announcements $\varphi!$ and $(\neg\varphi)!$: both remove links between φ-worlds and $\neg\varphi$-ones. The only difference is that the former can only take place at a current world which satisfies φ, and the latter in one satisfying $\neg\varphi$. This is reflected in valid principles of the logic, but we do not pursue this issue here.

Discussion: update and upgrade Distinguishing the two versions of information update also leads to a subtle distinction in a combined update-upgrade logic. If processing $!\varphi$ eliminates all worlds we know to be non-actual, our preference statements adjust automatically to what we know about the facts. This is the behavior of realists, who never cry over spilt milk. For those realists i, the following combined announcement/preference principle will be valid, at least for atomic statements p which do not change their truth values by being announced

$$[!p][pref]_i p.$$

But this principle is not valid for more nostalgic souls, who still deplore the way things turned out to be. For them, update amounts to the link-cutting operation $\varphi!$, they stick to their preferences between all possible worlds, and the new fact may even introduce regrets:

$$\langle pref\rangle_i \neg p \to [p!](\langle pref\rangle_i \neg p \wedge K_i p).$$

2.4 Dynamic epistemic upgrade logic

Language and semantics

Now we introduce an enriched dynamic language for update and upgrade. Its static part is the earlier language of Section 2.2, but its action vocabulary contains both link-cutting announcements $\varphi!$ and suggestions $\sharp\varphi$. Adding the original world-eliminating announcements $!\varphi$ is a routine matter, so we highlight the latter less standard variant only.

2.4.1. Definition. Let P be a set of proposition letters and I a set of agents, with p ranging over P, i over I. The *dynamic epistemic preference language* is given by the following rule:

$$\varphi ::= \bot \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_i\varphi \mid [pref]_i\varphi \mid U\varphi \mid [\pi]\varphi$$
$$\pi ::= \varphi! \mid \sharp\varphi.$$

We could also add the usual program operations of composition, choice, and iteration from propositional dynamic logic to the action vocabulary - but we have no special use for these in the current context. The new language can be interpreted on epistemic preference models as follows, where we choose the 'regret' variant of update for the novelty:

2.4.2. Definition. Given an epistemic preference model $\mathcal{M}$, the *truth definition* for formulas is as before, but with two new key clauses for the action modalities:

$(\mathcal{M}, s) \models [\varphi!]\psi$ iff if $\mathcal{M}, s \models \varphi$, then $\mathcal{M}_{\varphi!}, s \models \psi$.
$(\mathcal{M}, s) \models [\sharp\varphi]\psi$ iff $\mathcal{M}_{\sharp\varphi}, s \models \psi$.

Preference upgrade logic

On epistemic preference models, all valid principles of the static language of Section 2.2 still hold. Moreover, the usual axioms for public announcement hold, be it with one twist. As we saw, the usual updates $!\varphi$ eliminate all $\neg\varphi$-worlds, but updates $\varphi!$ leave all worlds in the model, cutting links instead. This makes no difference with purely epistemic dynamic axioms, but it does with global existential modalities over the whole domain of the model. The usual reduction axiom for operator E is this:

$$\langle !\varphi\rangle E\psi \leftrightarrow \varphi \wedge E\langle !\varphi\rangle\psi.$$

But the axiom below is different, as $E\varphi$ can still refer to worlds after the update which used to be $\neg\varphi$. Further comments will be found below. We focus on what is new here: upgrade, and its interplay with modified update. It is not hard to see the soundness of the following principles, stated with existential modalities for convenience:

2.4.3. Theorem. *The following formulas are valid:*

1. $\langle\varphi!\rangle p \leftrightarrow (\varphi \wedge p)$.
2. $\langle\varphi!\rangle\neg\psi \leftrightarrow (\varphi \wedge \neg\langle\varphi!\rangle\psi)$.
3. $\langle\varphi!\rangle(\psi \wedge \chi) \leftrightarrow (\langle\varphi!\rangle\psi \wedge \langle\varphi!\rangle\chi)$.
4. $\langle\varphi!\rangle\langle K\rangle_i\psi \leftrightarrow (\varphi \wedge \langle K\rangle_i\langle\varphi!\rangle\psi)$.
5. $\langle\varphi!\rangle\langle pref\rangle_i\psi \leftrightarrow (\varphi \wedge \langle pref\rangle_i\langle\varphi!\rangle\psi)$.
6. $\langle\varphi!\rangle E\psi \leftrightarrow (\varphi \wedge E(\langle\varphi!\rangle\psi \vee \langle\neg\varphi!\rangle\psi))$.
7. $\langle\sharp\varphi\rangle p \leftrightarrow p$.
8. $\langle\sharp\varphi\rangle\neg\psi \leftrightarrow \neg\langle\sharp\varphi\rangle\psi$.
9. $\langle\sharp\varphi\rangle(\psi \wedge \chi) \leftrightarrow (\langle\sharp\varphi\rangle\psi \wedge \langle\sharp\varphi\rangle\chi)$.
10. $\langle\sharp\varphi\rangle\langle K\rangle_i\psi \leftrightarrow \langle K\rangle_i\langle\sharp\varphi\rangle\psi$.
11. $\langle\sharp\varphi\rangle\langle pref\rangle_i\psi \leftrightarrow (\neg\varphi \wedge \langle pref\rangle_i\langle\sharp\varphi\rangle\psi) \vee (\langle pref\rangle_i(\varphi \wedge \langle\sharp\varphi\rangle\psi))$.
12. $\langle\sharp\varphi\rangle E\psi \leftrightarrow E\langle\sharp\varphi\rangle\psi$.

Proof. The first four formulas are the well-known valid reduction axioms for public announcement. The fifth formula, about commutation of $\langle\varphi!\rangle$ and $\langle pref\rangle_i$, expresses the fact that epistemic update does not change any preference relations. The special case of $E\varphi$ has been commented on above.

Next comes a similar set of reduction principles for upgrade. Axiom 7 is like Axiom 1, but simpler - as there is no precondition for $\sharp\varphi$: this operation can always be performed. Given that, we just state that atomic facts do not change under upgrade. The next two axioms express that upgrade is a function. Then comes a commutation principle for preference and knowledge which reflects the fact that upgrade does not change any epistemic relations.

Axiom 11 is crucial, as it encodes precisely how we change the preference relation. It says essentially this. After an upgrade for φ, a preference link leads from the current world to a φ-world if and only if this same link existed before. This means that it has not been removed, ruling out the case where it led from an actual world verifying φ to some other one verifying $\neg\varphi$. The three cases where the link does persist are described succinctly in the two disjuncts on the right-hand side. Finally, as the upgrade may have changed truth values of formulas, we must be careful, and say that, before the upgrade, the link went to a world satisfying $\langle\sharp\varphi\rangle$ rather than φ. The last axiom in the list is simply a commutativity principle for preference and existential modalities. □

This dynamic epistemic upgrade logic (henceforth, *DEUL*) can explain general effects of changes in information and preference. In particular, we can think of our upgrade system as transforming underlying *world-* or *object*-comparison relations, but then, in the matching logic, recording also what changes take place because of this at the level of *propositions*. Thus, given the earlier-noted expressive power of the modal language for notions of preference between propositions, we can derive principles telling us what new propositional preferences obtain after an upgrade action, and relate these to the propositional preferences that we had before. As an illustration, consider the $\forall\exists$-notion of preference stated earlier:

$$P^{\forall\exists}(\varphi,\psi) \quad \text{iff} \quad U(\psi \to \langle pref\rangle_i\varphi).$$

2.4.4. Fact. The following equivalence holds
$\langle\sharp A\rangle P^{\forall\exists}(\varphi,\psi)$ iff $P^{\forall\exists}(\langle\sharp A\rangle\varphi,\langle\sharp A\rangle\psi) \wedge P^{\forall\exists}((\langle\sharp A\rangle\varphi \wedge A),(\langle\sharp A\rangle\psi \wedge A))$.

Proof. This is a simple calculation showing how the dynamic epistemic upgrade logic axiom system works in practice:

$$\begin{aligned}
&\langle\sharp A\rangle P^{\forall\exists}(\varphi,\psi) \leftrightarrow \langle\sharp A\rangle U(\psi \to \langle pref\rangle_i\varphi)\\
&\leftrightarrow U(\langle\sharp A\rangle(\psi \to \langle pref\rangle_i\varphi))\\
&\leftrightarrow U(\langle\sharp A\rangle\psi \to \langle\sharp A\rangle\langle pref\rangle_i\varphi)\\
&\leftrightarrow U(\langle\sharp A\rangle\psi \to (\neg A \wedge \langle pref\rangle_i\langle\sharp A\rangle\varphi) \vee (\langle pref\rangle_i(A \wedge \langle\sharp A\rangle\varphi)))\\
&\leftrightarrow U(\langle\sharp A\rangle\psi \wedge \neg A \to \langle pref\rangle_i\langle\sharp A\rangle\varphi) \wedge U(\langle\sharp A\rangle\psi \wedge A \to\\
&\langle pref\rangle_i(\langle\sharp A\rangle\varphi \wedge A))\\
&\leftrightarrow P^{\forall\exists}(\langle\sharp A\rangle\varphi,\langle\sharp A\rangle\psi) \wedge P^{\forall\exists}((\langle\sharp A\rangle\varphi \wedge A),(\langle\sharp A\rangle\psi \wedge A)).
\end{aligned}$$

□

A similar analysis applies Von Wright's 'All All' notion of preference between propositions, relating new preferences in this sense to earlier ones – but we leave this calculation to the reader.[4]

In addition, as noted earlier, our epistemic upgrade logic can deal with combined scenarios like introducing 'regret'. Say, a sequence of instructions

$$\sharp p; \neg p! \quad \text{for atomic } p$$

will first make p attractive, and afterwards, unobtainable. The logic records this as the (derivable) validity of regret principles like that at the end of Section 2.3:

$$\langle pref \rangle_i p \to [\sharp p][\neg p!](\langle pref \rangle_i p \wedge K_i \neg p).$$

Dynamic epistemic upgrade logic can also analyze the basic propositional scenarios of obeying successive commands or reasoning toward achieving practical goals proposed in [Zar03] and [Yam06].

2.4.5. THEOREM. *Dynamic epistemic upgrade logic is completely axiomatized by the above reduction axioms.*

Proof. The reduction axioms, whose soundness we have already seen, are clearly sufficient for eventually turning every formula of our language into a static one without announcement or suggestion modalities. Then we can use the completeness theorem for our static language. □

The same reduction method also shows that *DEUL* is decidable.

We have reached the first major conclusion of this chapter:

Preference upgrade has a complete compositional logic-
just like, and even jointly with, knowledge update.

New issues of interest: coherence

Despite the technical analogies between information update and preference upgrade, there are also intuitive differences. One typical illustration is the intuitive notion of 'coherence'. In pure public announcement logics, the only relevant aspects of coherence for a sequence of assertions seem to be these:

(a) Do not make *inconsistent* and false assertions at the actual world; and, do not waste anyone's time.

[4] One might want to be more radical here, and insist on dynamic preference-changing actions directly *at the level of propositions*, without any dependence on an underlying world-level. This is in line with versions of belief revision theory where one is instructed to come to believe certain propositions. We have some thoughts on this alternative; but it would involve both entrenchment and preference relations on sets of propositions, a more syntactic perspective which raises as many design issues as the world-based semantic framework used in this chapter.

(b) Do not make assertions which are *common knowledge* in the whole group, and which do not change the model.

But in combination with upgrade, we can make other distinctions. E.g., the effect of a sequence with two conflicting suggestions

$$\sharp p; \sharp \neg p$$

is not inconsistency, but it still has some strange aspects. Generally speaking, such a sequence makes the ordering non-connected, as it removes arrows either way between p-worlds and $\neg p$-worlds. It is an interesting issue which sequences of upgrades are coherent, in that they preserve the property of connectedness.

In reality, one often resolves conflicts in suggestions by means of some authority ranking among the issuers of those suggestions. This is somewhat like the reality of information update. We often get contradictory information from different sources, and we need some notion of *reliability* differentiating between these to get to any sensible total update. Both issues go beyond the ambitions of this chapter, as they involve the gap between actual informational events and their translation into the idealized model changes offered by dynamic epistemic logics, whether for update or upgrade.

2.5 Relation change and product upgrade

Reduction axioms reflect definable operations

To a logician, standard epistemic update $!\varphi$ essentially relativizes a model $\mathcal{M}$ to a definable submodel $\mathcal{M}_{!\varphi}$. The relation between evaluation at both sides is expressed in the following standard result:

2.5.1. Fact. Assertions φ hold in the relativized model iff their syntactically relativized versions were true in the old model:

$$\mathcal{M}_{!\varphi} \models \psi \text{ iff } \mathcal{M} \models (\psi)^{\varphi}.$$

In this light, the reduction axioms for public announcement merely express the inductive facts about the modal assertion $\langle !\varphi \rangle \varphi$ referring to the left-hand side, relating these on the right to relativization instructions creating $(\psi)^{\varphi}$.

This same idea applies to preference upgrade $\sharp\varphi$. This time, the relevant semantic operation on models is *redefinition of base relations*. The same is true for the new link-cutting update operation $\varphi!$. [Ben07a] notes how relativization and redefinition make up the standard notion of *relative interpretation* between theories in logic when objects are kept fixed - while product update relates to more complex reductions forming new objects as tuples of old objects. In this light, the reduction axioms for *DEUL* reflect a simple inductive definition, this time

for what may be called *syntactic re-interpretation* of formulas. This operation leaves all logical operators unchanged, but it changes occurrences of the redefined relation symbol by its definition. There is one slight difference though. Relation symbols for preference only occur implicitly in our modal language, through the modalities. This is why the key reduction axiom in the above reflects a format of the following abstract recursive sort:

$$\langle R := def(R)\rangle\langle R\rangle\varphi \leftrightarrow \langle def(R)\rangle\langle R := def(R)\rangle\varphi.$$

Dynamic logic of relation changers

Further relation-changing operations can be defined, and make sense in our dynamic logics. We already mentioned the case of

$$R := R \cup (?\neg\varphi; \top; ?\varphi).$$

Here again, reduction axioms would be immediate, because of the following straightforward validities from propositional dynamic logic:

$$\begin{array}{lcl} \langle R \cup (?\neg\varphi; \top; ?\varphi)\rangle\psi & \leftrightarrow & \langle R\rangle\psi \vee \langle ?\neg\varphi; \top; ?\varphi\rangle\psi \\ & \leftrightarrow & \langle R\rangle\psi \vee (\neg\varphi \wedge E(\varphi \wedge \psi)). \end{array}$$

The example suggests a much more general observation, which we state informally in the following:

2.5.2. Fact. Every relation-changing operation that is definable in *PDL* without iteration has a complete set of reduction axioms in dynamic epistemic logic.

Proof. Clearly, every definition for a new relation $R^\sharp$ in this format is equivalent to a finite union of finite compositions of

(a) atomic relations R_i, (b) test relations $?\varphi$ for formulas of the base language. The standard *PDL* axioms for union, composition, and tests in *PDL* then rewrite all statements $\langle R^\sharp\rangle\varphi$ to compounds in terms of just basic modalities $\langle R_i\rangle\varphi$. □

This *PDL*-style analysis can even derive reduction axioms automatically:

2.5.3. Example. Our upgrade operation $\sharp\varphi$ is really the relation-changer:

$$R := (?\neg\varphi; R) \cup (R; ?\varphi).$$

Thus, the key reduction axiom can be derived as follows:

$$\begin{array}{lcl} \langle\sharp\varphi\rangle\langle R\rangle\psi & \leftrightarrow & \langle(?\neg\varphi; R) \cup (R; ?\varphi)\rangle\langle\sharp\varphi\rangle\psi \\ & \leftrightarrow & \langle ?\neg\varphi; R\rangle\langle\sharp\varphi\rangle\psi \vee \langle R; ?\varphi\rangle\langle\sharp\varphi\rangle\psi \\ & \leftrightarrow & (\neg\varphi \wedge \langle R\rangle\langle\sharp\varphi\rangle\psi) \vee \langle R\rangle(\varphi \wedge \langle\sharp\varphi\rangle\psi). \end{array}$$

The latter is just the version that we found 'by hand' in the above.

But we can do still better than this, and achieve the same generality as dynamic epistemic logics for information update – as will be shown briefly now.

Product update

The usual generalization of eliminative public announcement is product update ([Ger99], [BMS98], [DHK07]). We briefly recall the basics.

2.5.4. Definition. An *event model* is a tuple $\mathcal{E} = (E, \sim_i, PRE)$ such that E is a non-empty set of events, $\sim_i$ is a binary epistemic relation on E, PRE is a function from E to the collection of all epistemic propositions.

The intuition behind the function PRE is that it gives the *preconditions* for an action: an event a can be performed at world s only if the world a fulfills the precondition $PRE(a)$.

2.5.5. Definition. Given an *epistemic model* $\mathcal{M}$, an *event model* $\mathcal{E}$, the *product update model* $\mathcal{M} \times \mathcal{E}$ is defined as follows:

- The *domain* is $\{(s,\ a) \mid s$ a world in $\mathcal{M}$, a an event in $\mathcal{E}$, $(\mathcal{M}, s) \models PRE(a)\}$.
- The new *uncertainties* satisfy $(s, a) \sim_i (t, b)$ iff both $s \sim_i t$ and $a \sim_i b$.
- A world $(s,\ a)$ satisfies a *propositional atom* p iff s already did in $\mathcal{M}$.

2.5.6. Remark. For a version leaving all old worlds in place, as with the above new announcement operator $\varphi!$, we need to cut relational links again (instead of eliminating worlds) so as to 'isolate' those pairs (s, a) where $(\mathcal{M}, s)$ fails to satisfy the precondition for action a.

2.5.7. Definition. The language has new dynamic modalities $\langle \mathcal{E}, a \rangle$ referring to complex epistemic actions, and these are interpreted as follows:

$$\mathcal{M}, s \models \langle \mathcal{E}, a \rangle \varphi \text{ iff } \mathcal{M} \times \mathcal{E}, (s, a) \models \varphi.$$

This is the most powerful epistemic update calculus to date. As with public announcement, it yields a complete and decidable logic via a set of reduction axioms for all possible forms of postcondition (cf. [BMS98], [BEF93], [BEK06]).

Product upgrade

Next, we enrich epistemic event models with preference relations, indicating which events agents prefer over which others. These preferences may come from pay-offs or other benefits, but they may also be abstract relative plausibilities again, as in models of conditional logic.

2.5.8. Definition. The output for *product upgrade on epistemic preference models* are again the above epistemic models $\mathcal{M} \times \mathcal{E}$. But this time, we keep all world/action pairs (s, a) represented, as these are the non-realized options that we can still have regrets about. Then it remains to set the new preferences, and here, we can just follow the above direct product rule for relations:

$(s,t) \preceq_i (u,v)$ iff $s \preceq_i u$ and $t \preceq_i v$.

This product upgrade covers at least the earlier upgrade instruction $\sharp p$ for suggestions. To see this, consider the event model of Figure 2.4:

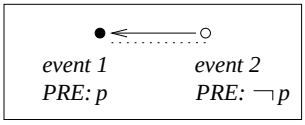

Figure 2.5: Two indistinguishable events

Here the two events cannot be distinguished epistemically by the agent. Recall that the reflexive loops of preference relations are omitted.

2.5.9. Fact. $\mathcal{M}_{\sharp\varphi} \cong \mathcal{M} \times \mathcal{E}^{\sharp\varphi}$, where the event model $\mathcal{E}^{\sharp\varphi}$ has two events "*seeing that* φ" (*event 1*), "*seeing that not-*φ" (*event 2*), with *event 2* $\preceq$ *event 1*.

Proof. From an epistemic viewpoint, the accessible part of $\mathcal{M} \times \mathcal{E}^{\sharp\varphi}$ merely *copies* the old model $\mathcal{M}$, as only one event can take place at each world. The old epistemic accessibilities just get copied with the product rule, since accessibility holds between all pairs of events. As for the new preference structure, consider any pair (s,t) in $\mathcal{M}$ where $\neg\varphi$ holds at s. Then the product model $\mathcal{M} \times \mathcal{E}^{\sharp\varphi}$ contains a unique corresponding pair

$$((s,\ \textit{event 2}), (t,\ \textit{event 1})).$$

Our product upgrade rule gives a preference here from left to right. The only case where this copying from $\mathcal{M}$ fails is when the old preference and the event preference do not match up. But this only happens in those cases where $\sharp\varphi$ would reject an existing link, namely, when $s \preceq t$, while $\mathcal{M}, s \models \varphi$ and $\mathcal{M}, t \models \neg\varphi$. □

Thus, as with public announcement and epistemic product update, one simple event model suffices to mimic our base mechanism for update or upgrade.

Much more generally, every upgrade rule which takes a current preference relation to a *PDL*-definable subrelation can be dealt with in the same style as above, by putting in enough events and preconditions. There are of course much more complex event models still, with many more worlds and complex preference relations for agents. These represent more refined scenarios for joint update and upgrade. We will return this issue later in Chapter 4.

Given the technical similarity of our product upgrade rule for preference to that for epistemic accessibility, the following is easy to see:

2.5.10. Theorem. *The dynamic logic of product update plus upgrade can be axiomatized completely by means of dynamic-epistemic-style reduction axioms.*

We are not going to spell out here what these axioms look like, but it is a routine exercise. Our second main conclusion in this chapter is this:

> *Preference upgrade can be combined naturally with the richest knowledge update mechanisms known so far.*

Virtues of the combination We think the above setting has independent interest. In philosophy, there is a well-known distinction between *preferences between states-of-affairs*, associated with 'consequentialist ethics', and *preferences between actions* in 'voluntarist ethics' (cf. [Sch97]). Our product update system models both kinds, and is able to study their interplay. Moreover, there is a computational angle, viz. 'dynamic deontic' versions of *PDL* itself, starting from preferences between worlds, but moving on to preferences between actions ([Mey88], [Mey96]). [PW04] follows up on the latter, and propose relation change as a way of 'changing policies'. [Roh05] provides a general background for this in so-called 'sabotage modal logic', where arbitrary links can be cut from models.

Thus, we see our product upgrade system also as one principled 'preferentialized' version of propositional dynamic logic.

2.6 Illustrations: defaults and obligations

We have presented an upgrade mechanism for incoming triggers that change preferences. We now illustrate this framework in two concrete settings. Our aim is not some full-fledged application to existing systems. We merely show how the logical issues in this chapter correspond to real questions of independent interest.

Default reasoning

Consider practical reasoning with default rules of the form "*if φ, then ψ*":

> "If I take the train right now, I will be home tonight".

These are defeasible conditionals, which recommend concluding ψ from φ, but without excluding the possibility of $\varphi \wedge \neg\psi$-worlds, be it that the latter are now considered exceptional circumstances. Intuitively, the latter are not 'ruled out' from our current model, but only 'downgraded' when a default rule is adopted. [Vel96] is an influential dynamic treatment, making a default an instruction for changing the current preference order between worlds. The simplest case has just one assertion φ which is being 'recommended' - in Veltman's terms, there is an instruction "*Normally, φ*". From our perspective, one can go this way, using a scenario of *relation change for defaults*, as in our earlier Section 2.3. Suppose that we want to give an incoming default rule "*Normally, φ*" 'priority', in that after its processing, all best worlds are indeed φ-worlds. Here is a more drastic procedure, which will validate the preceding intuition:

2.6.1. DEFINITION. We make all φ-worlds better than all $\neg\varphi$-worlds, and within the φ- and $\neg\varphi$-areas, we leave the old preferences in place.[5] Formally, this is one of our earlier *PDL*-style relation-changes: the old preference relation R becomes

$$(?\varphi; R; ?\varphi) \cup (?\neg\varphi; R; ?\neg\varphi) \cup (?\neg\varphi; \top; ?\varphi).$$

Interestingly, this is the union of the earlier link cutting version of public announcements $\varphi!$ plus the upgrade operation with relation extension considered in the preceding Section 2.4.

2.6.2. FACT. Relational default processing can be axiomatized completely.

Proof. By the method of Section 2.5, the key reduction axiom follows automatically from the given *PDL*-form, yielding

$$\langle\sharp\varphi\rangle\langle pref\rangle\psi \leftrightarrow\ (\varphi \wedge \langle pref\rangle(\varphi \wedge \langle\sharp\varphi\rangle\psi)$$
$$\vee\, (\neg\varphi \wedge \langle pref\rangle(\neg\varphi \wedge \langle\sharp\varphi\rangle\psi)) \vee (\neg\varphi \wedge E(\varphi \wedge \langle\sharp\varphi\rangle\psi)).$$ □

Thus, we have a plausible version of default logic in our upgrade setting. Moreover, their validities are axiomatizable in a systematic style via reduction axioms, rather than more ad-hoc default logics found in the literature.

Things need not stop here. E.g., the relation-changing version puts heavy emphasis on the last suggestion made, giving it the force of a command. This seems too strong in many cases, as it gears everything toward the last thing heard. A more reasonable scenario is this. We are given a sequence of instructions inducing preference changes, but they need not all be equally urgent. We need to find out our total commitments eventually. But the way we integrate these instructions may be partly left up to the *policy* that we choose, partly also to another parameter of the scenario: viz. the relative force or *authority* of the issuers of the instructions. One particular setting where this happens is again Optimality Theory. Ranked constraints determine the order of authority, but within that, one counts numbers of violations. Cf. [PS93] for a good exposition, and Chapter 3 for a logical exploration.

From default logic to belief revision Default logic is naturally connected with *belief revision*, since new facts may change earlier conclusions. More generally, an analysis of preference change seems very congenial to analyzing belief revision, with world ordering by relative plausibility (cf. [Gro88], [Rot06]). Indeed, the paper [Ben07a] shows that the techniques for handling relation change developed in this chapter can be used to analyze various belief revision policies, and axiomatize their properties completely.

[5] This is known as the 'lexicographic' change in the belief revision community. The idea was first suggested in [Nay94].

Deontic logic and commands

Similar considerations apply to deontic logic ([Åqv87]). Originally, this was the study of assertions of obligation

$O\varphi$: 'it ought to be the case that φ',

as well as statements of conditional obligation $O(\varphi|\psi)$, say, emanating from some moral authority. The sum total of all true O-statements represents all the obligations an agent has at the current stage.

In the standard semantics of deontic logic, $O\varphi$ is treated as a universal modality over some deontic accessibility relation. But the intuition is that those φ ought to be case which are true in *all best possible worlds*, as seen from the current one. Again, this suggests a preference order among worlds. And then, once more, we can think of this setting dynamically, using our upgrade scenario.

Initially, there are no preferences between worlds. Then some moral authority starts 'moralizing': introducing evaluative distinctions between worlds. If this process works well, we get a new ordering of worlds from which our current obligations may be computed, as those assertions which are true in all best worlds. Whether a sequence of commands makes sense in this way may depend on more than consistency, and the issue of 'coherence' in Section 2.3 comes back again with greater force now.

Looking backward, or forward in upgrade Deontic logic also raises new issues. One semantic intuition is that, after a command (say, 'Thou shalt not kill'), the core proposition becomes true in all best possible worlds. Thus, in commands, there is a future-oriented aspect:

'See to it that φ' should result in a new situation where $O\varphi$ is true.

But as we have seen in Section 2.4, not every upgrade $\sharp\varphi$ has the effect that φ becomes true in the new most preferred worlds. Indeed, there is a general difficulty with specifications of the form 'See to it that φ'. DEL is mainly about events with their preconditions. Thus, the information one gets from an event is *past-oriented*, describing what was the case at the time the event happened. But, even a simple epistemic event can change the truth value of assertions at worlds - witness public announcements turning ignorance into knowledge.

But it is not so easy to just define an action as achieving the truth of some proposition. This works for simple factual effects of actions like opening a door ([BOR06]), but it is not clear what this should even mean with more complex stipulations. E.g., there is no obvious 'seeing to it that' arbitrary mixtures of knowledge and ignorance in groups arise, and the same seems true of complex deontic commands. Whether deontic reasoning needs some sort of future-oriented update and upgrade seems an interesting question. For temporal logics of such $STIT$ operators, cf. [BPX01].

2.7 Related work

The ideas in this chapter have a long history, and there are many proposals in the literature having to do with 'dynamification' of preferences, defaults, and obligations. We just mention a few related approaches here, though we do not make any detailed comparisons. [Mey88] was probably the first to look at deontic logic from a dynamic point of view, with the result that deontic logics are reduced to suitable versions of dynamic logics. This connection has become a high-light in computer science since, witness the regular *DEON* conference series. In a line that goes back to [Spo88], [Vel96] presents an update semantics for default rules, locating their meaning in the way in which they modify expectation patterns. This is part of the general program of 'update semantics' for conditionals and other key expressions in natural language. [TT99] use ideas from update semantics to formalize deontic reasoning about obligations, but with motivations from computer science. In their view, the meaning of a normative sentence resides in the changes it brings about in the 'ideality relations' of agents to whom the norm applies. [Mey96] takes the deontic logic/dynamic logic interface a step further, distinguishing two notions of permission, one of which, 'free choice permission' requires a new 'dynamic logic of permission', where preferences can hold between actions. Completeness theorems with respect to this enriched semantics are given for several systems. Taking belief change as its starting point, [Han95] identified four types of changes in preference, namely revision, contraction, addition and subtraction, and showed that they satisfy plausible postulates for rational changes in preferences. [PW04] provide a dynamified version of the dynamic logic of permission, in order to deal with building up of agents' policies by adding or deleting transitions. [Dem05] reduces an extension of van der Meyden's logic to propositional dynamic logic, yielding an EXPTIME decision procedure, and showing how dynamic logic can deal with agents' policies. Following van Benthem's 'sabotage games', [Roh05] studies general modal logics with operators that describe effects of deleting arbitrary transitions - without a fixed upgrade definition as in our analysis. Model checking for such logics becomes PSPACE-complete, and satisfiability is undecidable. [PPC06] observe that an agent's obligations are often dependent on what she knows, and introduce a close relative of our epistemic preference language, but over temporal tree models. They provide distinctions, like knowing one's duty versus having a duty to know, whose dynamics invites a merge with our system. Our own approach goes back to [BEF93], which discusses general formats for upgrading preference relations. [Zar03] uses similar ideas, combined with a simple update logic to formalize natural language imperatives of the form $FIAT\ \varphi$, which can be used in describing the search for solutions of given planning problems. More generally, [Yam06] takes the update paradigm to logics of commands and obligations, modeling changes brought about by various acts of commanding. It combines a multi-agent variant of the language of monadic deontic logic with a dynamic language for updates and commands. This

is closest to what we do. Yamada's command operator for propositions A can be modeled *exactly* as an upgrade sending R to $R; ?A$ in our system. But this chapter provides a much more general treatment of possible upgrade instructions. Finally, [Rot06] presents a format for relation change which can handle all major current policies for belief revision. [Ben07a] shows how one can axiomatize such policies completely using the methods in Section 2.5 of this chapter.

A full-fledged comparison doing justice to all these approaches is unfortunately beyond the scope of this chapter.

2.8 Conclusion

In this chapter we have shown that preference upgrade is a natural and crucial part of logical dynamics. It can be modeled essentially as relation change in a standard dynamic format, up to the expressive level of the best available system, that of epistemic product update.

Still, our approach leaves things to be desired. In particular, many settings call for more finely-grained distinctions as to *intensity* of preferences, as happens in quantitative versions of social choice theory. [Liu06b] has proposed a mechanism of *utility update*, inspired by [Spo88], [Auc03] and [Liu04], which combines utilities of old worlds and of events to compute utilities of new worlds. With such a system, we can upgrade defaults, duties, or preferences in games in a more controlled local fashion, by adding or subtracting 'points'. The relationship between our relational upgrade and more quantitative utility update also poses some interesting technical issues, for which we refer to the more extensive exploration in [Liu06b].

Chapter 3

Preference, Priorities and Belief

3.1 Motivation

The notion of preference occurs frequently in game theory, decision theory, and many other research areas. Typically, preference is used to draw comparison between two alternatives explicitly. Studying preference and its general properties has become a main logical concern after the pioneering seminar work by [Hal57] and [Wri63], witness [Jen67], [Cre71], [Tra85], [DW94], [Han01a], [BRG07] etc., and more recently work on dynamics of preference e.g. [Han95] and [BL07]. Let us single out immediately the two distinctive characteristics of the approach to preference we take in this chapter.

- Most of the previous work has taken preference to be a primitive notion, without considering how it comes into being. We take a different angle here and explore both preference and its origin. We think that preference can often be rationally derived from a more basic source, which we will call a *priority base*. In this manner we have two levels: the priority base, and the preference derived from it. We hope this new perspective will shed light on the reasoning underlying preference, so that we are able to discuss *why* we prefer one thing over another. There are many ways to get preference from such a priority base, a good overview can be found in [CMLLM04].

- In real life we often encounter situations in which no complete information is available. Preference will then have to be based on our beliefs, i.e. do we believe certain properties from the priority base to apply or not? Apparently, this calls for a combination of doxastic language and preference language. We will show a close relationship between preference and beliefs. To us, both are mental attitudes. If we prefer something, we believe we do (and conversely). In addition, this chapter is also concerned with the dynamics of preference. By means of our approach, we can study preference

changes, whether they are due to a change in the priority base, or caused by belief revision.

Depending on the actual situation, preference can be employed to compare alternative states of affairs, objects, actions, means, and so on, as listed in [Wri63]. One requirement we impose is that we consider only mutually exclusive alternatives. In this paper, we consider in first instance preference over *objects* rather than between propositions (compare [DW94]). Objects are, of course, congenitally mutually exclusive. Although the priority base approach is particularly well suited to compare preference between objects, it can be applied to the study of the comparison of other types of alternatives as well. In Section 3.7 we show how to apply the priority base approach to propositions. When comparing objects, the kind of situation to be thought of is:

3.1.1. Example. Alice is going to buy a house. For her there are several things to consider: the cost, the quality and the neighborhood, strictly in that order. All these are clear-cut for her, for instance, the cost is good if it is inside her budget, otherwise it is bad. Her decision is then determined by the information whether the alternatives have the desirable properties, and by the given order of importance of the properties.

In other words, Alice's preference regarding houses is derived from the priority order of the properties she considers. This chapter aims to propose a logic to model such situations. When covering situations in which Alice's preference is based on incomplete information belief will enter into the logic as an operation.

There are several points to be stressed beforehand, in order to avoid misunderstandings: First, our intuition of priority base is linked to graded semantics, e.g. spheres semantics by [Lew73]. We take a rather syntactical approach in this chapter, but that is largely a question of taste, one can go about it semantically as well. We will return to this point several times. Second, we will mostly consider a linearly ordered priority base. This is simple, giving us a quasi-linear order of preference. But our approach can be adapted to the partially ordered case, as we will indicate at the end of the paper. Third, when we add a belief operator to the preference language (fragment of FOL), it may seem that we are heading into doxastic predicate logic. This is true, but we are not going to be affected by the existing difficult issues in that logic. What we are using in this context is a very limited part of the language. Finally, although we start with a two level perspective this results on the preference side in logics that are rather like ordinary propositional modal logics. The bridge between the two levels is then given by theorems that show that any models of these modal logics can be seen as having been constructed from a priority base. These theorems are a kind of completeness theorems, but we call them *representation theorems* to distinguish them from the purely modal completeness results.

The following sections are structured as follows: In Section 3.2, we start with a simple language to study the rigid case in which the priorities lead to a clear and unambiguous preference ordering. In Section 3.3 we review some basics about ordering. Furthermore, a proof of a representation theorem for the simple language without beliefs is presented. Section 3.4 will consider what happens when the agent has incomplete information about the priorities with regard to the alternatives. In Section 3.5 we will look at changes in preference caused by two different sources: changes in beliefs, and changes of the sequence of priorities. Section 3.6 is an extension to the multi-agent system. We will prove representation theorems for the general case, and for the special cases of cooperative agents and competitive agents. In Section 3.7 we apply our approach to preference over propositions. Finally, we discuss how to generalize our approach to partially ordered preferences, and we end the chapter with a few conclusions.

3.2 From priorities to preference

As we mentioned in the preceding, there are many ways to derive preference from the priority base. We choose one of the mechanisms, the way of Optimality Theory (OT), as an illustration because we like the intuition behind this mechanism. Along the way, we will discuss other approaches as well, to indicate how our method can be applied to them as well.

Here is a brief review of some ideas from optimality theory that are relevant to the current context. In optimality theory a set of conditions is applied to the alternatives generated by the grammatical or phonological theory, to produce an optimal solution. It is by no means sure that the optimal solution satisfies all the conditions. There may be no such alternative. The conditions, called *constraints*, are strictly ordered according to their importance, and the alternative that satisfies the earlier conditions best (in a way described more precisely below) is considered to be the optimal one. This way of choosing the optimal alternative naturally induces a preference ordering among all the alternatives. We are interested in formally studying the way the constraints induce the *preference ordering* among the alternatives. The attitude in our investigations is somewhat differently directed than in optimality theory.[1]

Back to the issues of preference, to discuss preference over objects, we use a first order logic with constants $d_0, d_1 \ldots$; variables $x_0, x_1, \ldots$; and predicates $P, Q, P_0, P_1, \ldots$. In practice, we are thinking of finite domains, monadic predi-

[1]Note that in optimality theory the optimal alternative is chosen unconsciously; we are thinking mostly of applications where conscious choices are made. Also, in optimality theory the application of the constraints to the alternatives lead to a *clear* and *unambiguous* result: either the constraint clearly is true of the alternative or it is not, and that is something that is not sensitive to change. We will loosen this condition and consider issues that arise when changes do occur.

cates, simple formulas, usually quantifier free or even variable free. The following definition is directly inspired by optimality theory, but to take a neutral stance we use the words priority sequence instead of constraint sequence.

3.2.1. Definition. A *priority sequence* is a finite ordered sequence of formulas (priorities) written as follows:

$$C_1 \gg C_2 \cdots \gg C_n \quad (n \in \mathbb{N}),$$

where each of C_m $(1 \leq m \leq n)$ is a formula from the language, and there is exactly one free variable x, which is a common one to each C_m.

We will use symbols like $\mathfrak{C}$ to denote priority sequences. The priority sequence is linearly ordered. It is to be read in such a way that the earlier priorities count strictly heavier than the later ones, for example, $C_1 \wedge \neg C_2 \wedge \cdots \wedge \neg C_m$ is preferable over $\neg C_1 \wedge C_2 \wedge \cdots \wedge C_m$ and $C_1 \wedge C_2 \wedge C_3 \wedge \neg C_4 \wedge \neg C_5$ is preferable over $C_1 \wedge C_2 \wedge \neg C_3 \wedge C_4 \wedge C_5$. A difference with optimality theory is that we look at *satisfaction* of the priorities whereas in optimality theory *infractions* of the constraints are stressed. This is more a psychological than a formal difference. However, optimality theory knows multiple infractions of the constraints and then counts the number of these infractions. We do not obtain this with our simple objects, but we think that possibility can be achieved by considering composite objects, like strings.

3.2.2. Definition. Given a priority sequence of length n, two objects x and y, *Pref(x,y)* is defined as follows:

$$Pref_1(x,y) ::= C_1(x) \wedge \neg C_1(y),$$
$$Pref_{k+1}(x,y) ::= Pref_k(x,y) \vee (Eq_k(x,y) \wedge C_{k+1}(x) \wedge \neg C_{k+1}(y)), k < n,$$
$$Pref(x,y) ::= Pref_n(x,y),$$

where the auxiliary binary predicate $Eq_k(x,y)$ stands for $(C_1(x) \leftrightarrow C_1(y)) \wedge \cdots \wedge (C_k(x) \leftrightarrow C_k(y))$.[2]

In Example 3.1.1, Alice has the following priority sequence:

$$C(x) \gg Q(x) \gg N(x),$$

where $C(x)$, $Q(x)$ and $N(x)$ are intended to mean 'x has low cost', 'x is of good quality' and 'x has a nice neighborhood', respectively. Consider two houses d_1 and d_2 with the following properties: $C(d_1), C(d_2), \neg Q(d_1), \neg Q(d_2), N(d_1)$ and $\neg N(d_2)$. According to the definition, Alice prefers d_1 over d_2, i.e. $Pref(d_1, d_2)$.

Unlike in Section 3.4 belief does not enter into this definition. This means that $Pref(x,y)$ can be read as *x is superior to y*, or *under complete information x is preferable over y*.

[2]This way of deriving an ordering from a priority sequence is called *leximin ordering* in [CMLLM04].

3.2.3. Remark. Our method easily applies when the priorities become graded. Take the Example 3.1.1, if Alice is more particular, she may split the cost C into C^1 very low cost, C^2 low cost, C^3 medium cost, similarly for the other priorities. The original priority sequence $C(x) \gg Q(x) \gg N(x)$ may change into

$$C^1(x) \gg C^2(x) \gg Q^1(x) \gg C^3(x) \gg Q^2(x) \gg N^1(x) \gg \dots \ .$$

As we mentioned at the beginning, we have chosen a syntactic approach expressing priorities by formulas. If we switch to a semantical point of view, the priority sequence translates into pointing out a sequence of n sets in the model. The elements of the model will be objects rather than worlds as is usual in this kind of study. But one should see this really as an insignificant difference. If one prefers, one may for instance in Example 3.1.1 replace house d by the situation in which Alice has bought the house d.

When one points out sets in a model, Lewis' sphere semantics ([Lew73] p.98-99) comes to mind immediately. The n sets in the model obtained from the priority base are in principle unrelated. In the sphere semantics the sets which are pointed out are linearly ordered by inclusion. To compare with the priority base we switch to a syntactical variant of sphere semantics, a sequence of formulas $G_1, \dots, G_m$ such that $G_i(x)$ implies $G_j(x)$ if $i \leq j$. These formulas express the preferability in a more direct way, $G_1(x)$ is the most preferable, $G_m(x)$ the least. In what follows, we will show that the two approaches are equivalent in the sense that they can be translated into each other.

3.2.4. Theorem. *A priority sequence $C_1 \gg C_2 \cdots \gg C_m$ gives rise to a G-sequence of length 2^m. In the other direction a priority sequence can be obtained from a G-sequence logarithmic in the length of the G-sequence.*

Proof. Let us just look at the case that m=3. Assuming that we have the priority sequence $C_1 \gg C_2 \gg C_3$, the preference of objects is decided by where their properties occur in the following list:

$R_1 : C_1 \wedge C_2 \wedge C_3$;
$R_2 : C_1 \wedge C_2 \wedge \neg C_3$;
$R_3 : C_1 \wedge \neg C_2 \wedge C_3$;
$R_4 : C_1 \wedge \neg C_2 \wedge \neg C_3$;
$R_5 : \neg C_1 \wedge C_2 \wedge C_3$;
$R_6 : \neg C_1 \wedge C_2 \wedge \neg C_3$;
$R_7 : \neg C_1 \wedge \neg C_2 \wedge C_3$;
$R_8 : \neg C_1 \wedge \neg C_2 \wedge \neg C_3$.

The $G_i s$ are constructed as disjunctions of members of this list. In their most simple form, they can be stated as follows:

$G_1 : R_1$;
$G_2 : R_1 \vee R_2$;
$\vdots$
$G_8 : R_1 \vee \ R_2 \cdots \vee R_8$.

On the other hand, given a G_i-sequence, we can define C_i as follows,

$C_1 = R_1 \vee R_2 \vee R_3 \vee R_4$;
$C_2 = R_1 \vee R_2 \vee R_5 \vee R_6$;
$C_3 = R_1 \vee R_3 \vee R_5 \vee R_7$.

And again this can be simply read off from a picture of the G-spheres. The relationship between C_i, R_i, and G_i can be seen from the Figure 3.1. □

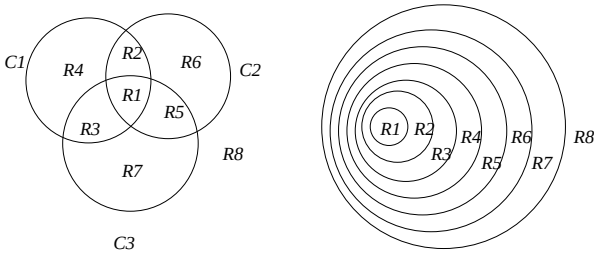

Figure 3.1: C_i, R_i, and G_i

3.2.5. Remark. In applying our method to such spheres, the definition of $\underline{Pref}(x, y)$ comes out to be $\forall i(y \in G_i \rightarrow x \in G_i)$. The whole discussion implies of course that our method can be applied to spheres as well as to any other approach which can be reduced to spheres.

3.2.6. Remark. As we pointed out at the beginning, one can define preference from a priority sequence $\mathfrak{C}$ in various different ways, all of which we can handle. Here is one of these ways, called *best-out ordering* in [CMLLM04], as an illustration. We define the preference as follows:

$$Pref(x, y) \quad \text{iff} \quad \exists C_j \in \mathfrak{C}(\forall C_i \gg C_j \, ((C_i(x) \wedge C_i(y)) \wedge (C_j(x) \wedge \neg C_j(y))).$$

In this case, we only continue along the priority sequence as long as we receive positive information. Returning the Example 3.1.1, this means that under this option we only get the conclusion that $\underline{Pref}(d_1, d_2)$ and $\underline{Pref}(d_2, d_1)$: d_1 and d_2 are equally preferable, because after observing that $\neg Q(d_1), \neg Q(d_2)$, Alice won't consider N at all.

3.3 Order and a representation theorem

In this section we will just run through the types of order that we will use in the current context. A relation $<$ is a *linear order* if $<$ is irreflexive, transitive and asymmetric, and satisfies *connectedness*:

$$x < y \vee x = y \vee y < x$$

More precisely, $<$ is called a *strict* linear order. A *non-strict* linear order $\leq$ is a reflexive, transitive, antisymmetric and connected relation. It is for various reasons useful to introduce non-strict variants of orderings as well.

Mathematically, strict and non-strict linear orders can easily be translated into each other:

(1) $x < y \leftrightarrow x \leq y \wedge x \neq y$, or
(2) $x < y \leftrightarrow x \leq y \wedge \neg(y \leq x)$,
(3) $x \leq y \leftrightarrow x < y \vee x = y$, or
(4) $x \leq y \leftrightarrow x < y \vee (\neg(x < y) \wedge \neg(y < x))$.

Optimality theory only considers linearly ordered constraints. These will be seen to lead to a *quasi-linear order* of preferences, i.e. a relation $\preccurlyeq$ that satisfies all the requirements of a non-strict linear order but antisymmetry. A quasi-linear ordering contains *clusters* of elements that are 'equally large'. Such elements are $\leq$ each other. Most naturally one would take for the strict variant $\prec$ an irreflexive, transitive, connected relation. If one does that, strict and non-strict orderings can still be translated into each other (only by using alternatives (2) and (4) in the above though, not (1) and (3)). However, $Pref$ is normally taken to be an asymmetric relation, and we agree with that, so we take the option of $\prec$ as an irreflexive, transitive, asymmetric relation. Then $\prec$ is definable in terms of $\preccurlyeq$ by use of (2), but not $\preccurlyeq$ in terms of $\prec$. That is clear from the picture below, an irreflexive, transitive, asymmetric relation cannot distinguish between the two given orderings.

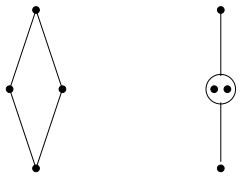

Figure 3.2: Incomparability and indifference.

One needs an additional equivalence relation $x \sim y$ to express that x and y are elements in the same cluster; $x \sim y$ can be defined by

(5) $x \sim y \leftrightarrow x \leq y \wedge y \leq x$.

Then, in the other direction, $x \leq y$ can be defined in terms of $<$ and $\sim$:

(6) $x \leq y \leftrightarrow x < y \vee x \sim y$.

It is certainly possible to extend our discussion to partially ordered sets of constraints, and we will make this excursion in Section 3.8. The preference relation will no longer be a quasi-linear order, but a so-called *quasi-order*: in the non-strict case a reflexive and transitive relation, in the strict case an asymmetric, transitive relation. One can still use (2) to obtain a strict quasi-order from a non-strict one and (6) to obtain a non-strict quasi-order from a strict one and $\sim$. However, we will see in Section 3.4 that in some contexts involving beliefs these translations no longer give the intended result. In such a case one has to be satisfied with the fact that (5) still holds and that $\prec$ as well as $\sim$ imply $\preccurlyeq$.

In the following we will write $Pref$ for the strict version of preference, $\underline{Pref}$ for the non-strict version, and let Eq correspond to $\sim$, expressing two elements are equivalent. Clearly, no matter what the priorities are, the non-strict preference relation has the following general properties:

(a) $\underline{Pref}(x, x)$,
(b) $\underline{Pref}(x, y) \vee \underline{Pref}(y, x)$,
(c) $\underline{Pref}(x, y) \wedge \underline{Pref}(y, z) \rightarrow \underline{Pref}(x, z)$.

(a), (b) and (c) express reflexivity, connectedness and transitivity, respectively. Thus, $\underline{Pref}$ is a quasi-linear relation; it lacks antisymmetry.

Unsurprisingly, (a), (b) and (c) are a complete set of principles for preference. We will put this in the form of a representation theorem as we announced in the introduction. In this case it is a rather trivial matter, but it is worthwhile to execute it completely as an introduction to the later variants. We reduce the first order language for preference to its core:

3.3.1. Definition. Let Γ be a set of propositional variables, and D be a finite domain of objects, the *reduced language* of preference logic is defined as follows,

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \underline{Pref}(d_i, d_j),$$

where p, d_i respectively denote elements from Γ and D.

The reduced language contains the propositional calculus. From this point onwards we refer to the language with variables, quantifiers, predicates as the *extended language*. In the reduced language, we rewrite the axioms as follows:

(a) $\underline{Pref}(d_i, d_i)$,
(b) $\underline{Pref}(d_i, d_j) \vee \underline{Pref}(d_j, d_i)$,
(c) $\underline{Pref}(d_i, d_j) \wedge \underline{Pref}(d_j, d_k) \rightarrow \underline{Pref}(d_i, d_k)$.

We call this axiom system **P**.

3.3.2. THEOREM. (*representation theorem*). $\vdash_{\mathbf{P}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences.*

Proof. The direction from left to right is obvious. Assume formula $\varphi(d_1, \ldots, d_n, p_1, \ldots, p_k)$ is not derivable in **P**. Then a non-strict quasi-linear ordering of the $d_1, \ldots, d_n$ exists, which, together with a valuation of the atoms $p_1, \ldots, p_k$ in φ falsifies $\varphi(d_1, \ldots, d_n)$. Let us assume that we have a linear order (adaptation to the more general case of quasi-linear order is simple), and also, w.l.o.g. that the ordering is $d_1 > d_2 > \cdots > d_n$. Then we introduce an extended language containing unary predicates $P_1, \ldots, P_n$ with a priority sequence $P_1 \gg P_2 \cdots \gg P_n$ and let P_i apply to d_i only. Clearly, the preference order of $d_1, \ldots, d_n$ with respect to the given priority sequence is from left to right. We have transformed the model into one in which the defined preference has the required properties.[3] □

3.3.3. REMARK. It is instructive to execute the above proof for the reduced language containing some additional predicates $Q_1, \ldots, Q_k$. One would like then to obtain a priority sequence of formulas in the language built up from Q_1 to Q_k. This is possible if in the model $\mathcal{M}$ each pair of constants d_i and d_j is distinguishable by formulas in this language, i.e. for each i and j, there exists a formula φ_{ij} such that $\mathcal{M} \models \varphi_{ij}(d_i)$and $\mathcal{M} \models \neg\varphi_{ij}(d_j)$. In such a case, the formula $\psi_i = \bigwedge_{i \neq j} \varphi_{ij}$ satisfies only d_i. And $\psi_1 \gg \cdots \gg \psi_n$ is the priority sequence as required. It is necessary to introduce new predicates when two constants are indistinguishable. A trivial method to do this is to allow identity in the language, $x = d_1$ obviously distinguishes d_1 and d_2.

Let us at this point stress once more what the content of a representation theorem is. It tells us that the way we have obtained the preference relations, namely from a priority sequence, does not affect the general reasoning about preference, its logic. The above proof shows this in a rather strong way: if we have a model in which the preference relation behaves in a certain manner, then we can think of this preference as derived from a priority sequence without disturbing the model as it is.

3.4 Preference and belief

In this section, we discuss the situation that arises when an agent has only incomplete information, but she likes to express her preference. The language will be extended with belief operators $B\varphi$ to deal with such uncertainty, and it is a

[3]Note that, although we used n priorities in the proof to make the procedure easy to describe, in general $^2log(n) + 1$ priorities are sufficient for the purpose.

small fragment of doxastic predicate logic. It would be interesting to consider what more the full doxastic predicate logic language can bring us, but we will leave this question to other occasions. We will take the standard **KD45** as the logic for beliefs, though we are aware of the philosophical discussions on beliefs and the options of proper logical systems.

Interestingly, the different definitions of preference we propose in the following spell out different "procedures" an agent may follow to decide her preference when processing the incomplete information about the relevant properties. Which procedure is taken strongly depends on the domain or the type of agents. In the new language, the definition of priority sequence remains the same, i.e. a priority C_i is a formula from the language *without* belief operators.

3.4.1. Definition. (decisive preference). Given a priority sequence of length n, two objects x and y, *Pref(x,y)* is defined as follows:

$Pref_1(x,y) ::= BC_1(x) \wedge \neg BC_1(y)$,
$Pref_{k+1}(x,y) ::= Pref_k(x,y) \vee (Eq_k(x,y) \wedge BC_{k+1}(x) \wedge \neg BC_{k+1}(y)), k < n$,
$Pref(x,y) ::= Pref_n(x,y)$,

where $Eq_k(x,y)$ stands for $(BC_1(x) \leftrightarrow BC_1(y)) \wedge \cdots \wedge (BC_k(x) \leftrightarrow BC_k(y))$.

To determine the preference relation, one just runs through the sequence of relevant properties to check whether one believes them of the objects. But at least two other options of defining preference seem reasonable as well.

3.4.2. Definition. (conservative preference). Given a priority sequence of length n, two objects x and y, *Pref(x,y)* is defined below:

$Pref_1(x,y) ::= BC_1(x) \wedge B\neg C_1(y)$,
$Pref_{k+1}(x,y) ::= Pref_k(x,y) \vee (Eq_k(x,y) \wedge BC_{k+1}(x) \wedge B\neg C_{k+1}(y)), k < n$,
$Pref(x,y) ::= Pref_n(x,y)$

where $Eq_k(x,y)$ stands for $(BC_1(x) \leftrightarrow BC_1(y)) \wedge (B\neg C_1(x) \leftrightarrow B\neg C_1(y)) \wedge \cdots \wedge$
$(BC_k(x) \leftrightarrow BC_k(y)) \wedge (B\neg C_k(x) \leftrightarrow B\neg C_k(y))$.

3.4.3. Definition. (deliberate preference). Given a priority sequence of length n, two objects x and y, *Pref(x,y)* is defined below:

$Supe_1(x,y)$[4] $::= C_1(x) \wedge \neg C_1(y)$,
$Supe_{k+1}(x,y) ::= Supe_k(x,y) \vee (Eq_k(x,y) \wedge C_{k+1}(x) \wedge \neg C_{k+1}(y)), k < n$,
$Supe(x,y) ::= Supe_n(x,y)$,

[4]Superiority is just defined as preference was in the previous section.

$Pref(x, y) ::= B(Supe(x, y))$,

where $Eq_k(x, y)$ stands for $(C_1(x) \leftrightarrow C_1(y)) \wedge \cdots \wedge (C_k(x) \leftrightarrow C_k(y))$.

To better understand the difference between the above three definitions, we look at the Example 3.1.1 again, but in three different variations:

A. Alice favors Definition 3.4.1: She looks at what information she can get, she reads that d_1 has low cost, about d_2 there is no information. This immediately makes her decide for d_1. This will remains so, no matter what she hears about quality or neighborhood.

B. Bob favors Definition 3.4.2: The same thing happens to him. But he reacts differently than Alice. He has no preference, and that will remain so as long as he hears nothing about the cost of d_2, no matter what he hears about quality or neighborhood.

C. Cora favors Definition 3.4.3: She also has the same information. On that basis Cora cannot decide either. But some more information about quality and neighborhood helps her to decide. For instance, suppose she hears that d_1 has good quality or is in a good neighborhood, and d_2 is not of good quality and not in a good neighborhood. Then Cora believes that, no matter what, d_1 is superior, so d_1 is her preference. Note that such kind of information could not help Bob to decide.

Speaking more generally in terms of the behaviors of the above agents, it seems that Alice always decides what she prefers on the basis of the limited information she has. In contrast, Bob chooses to wait and require more information. Cora behaves somewhat differently, she first tries to do some reasoning with all the available information before making her decision. This suggests yet another perspective on diversity of agents than discussed in Chapter 6.

Apparently, we have the following fact.

3.4.4. Fact.

- Totality holds for Definition 3.4.1, but not for Definition 3.4.2 or 3.4.3;

- Among the above three definitions, Definition 3.4.2 is the strongest in the sense that if $Pref(x, y)$ holds according to Definition 3.4.2, then $Pref(x, y)$ holds according to Definition 3.4.1 and 3.4.3 as well.

It is striking that, if in Definition 3.4.3, one plausibly also defines $\underline{Pref}(x, y)$ as $B(\underline{Supe}(x, y))$, then the normal relation between $Pref$ and $\underline{Pref}$ no longer holds: $Pref$ is not definable in terms of $\underline{Pref}$ any more, or even $\underline{Pref}$ in terms of $Pref$ and Eq.

For all three definitions, we have the following theorem.

3.4.5. THEOREM. $\underline{Pref}(x,y) \leftrightarrow B\underline{Pref}(x,y)$.

Proof. In fact we prove something more general in **KD45**. Namely, if α is a propositional combination of B-statements, then $\vdash_{\mathbf{KD45}} \alpha \leftrightarrow B\alpha$.

From left to right, since α is a propositional combination of B-statements, it can be transformed into conjunctive normal form: $\beta_1 \vee \cdots \vee \beta_k$. It is clear that $\vdash_{\mathbf{KD45}} \beta_i \rightarrow B\beta_i$ for each i, because each member γ of the conjunction β_i implies $B\gamma$. If $A = \beta_1 \vee \cdots \vee \beta_k$ holds then some β_i holds, so $B\beta_i$, so $B\alpha$. Then we immediately have: $\vdash_{\mathbf{KD45}} \neg\alpha \rightarrow B\neg\alpha$ (*) as well, since $\neg\alpha$ is also a propositional combination of B-statements if α is.

From right to left: Suppose $B\alpha$ and $\neg\alpha$. Then $B\neg\alpha$ by (*), so $B\bot$, but this is impossible in **KD45**, therefore α holds.

The theorem follows since $\underline{Pref}(x,y)$ is in all three cases indeed a propositional combination of B-statements. □

3.4.6. COROLLARY. $\neg\underline{Pref}(x,y) \leftrightarrow B\neg\underline{Pref}(x,y)$.

Actually, we think it is proper that Theorem 3.4.5 and Corollary 3.4.6 hold because we believe that preference describes a state of mind in the same way that belief does. Just as one believes what one believes, one believes what one prefers.

If we stick to Definition 3.4.1, we can generalize the representation result (Theorem 3.3.2). Let us consider the reduced language built up from standard propositional letters, plus $\underline{Pref}(d_i,d_j)$ by the connectives, and belief operators B. Again we have the normal principles of **KD45** for B.

3.4.7. THEOREM. *The following principles axiomatize exactly the valid ones.*

(*a*) $\underline{Pref}(d_i,d_i)$,
(*b*) $\underline{Pref}(d_i,d_j) \vee \underline{Pref}(d_j,d_i)$,
(*c*) $\underline{Pref}(d_i,d_j) \wedge \underline{Pref}(d_j,d_k) \rightarrow \underline{Pref}(d_i,d_k)$,
(*1.*) $\neg B\bot$,
(*2.*) $B\varphi \rightarrow BB\varphi$,
(*3.*) $\neg B\varphi \rightarrow B\neg B\varphi$,
(*4.*) $\underline{Pref}(d_i,d_j) \leftrightarrow B\underline{Pref}(d_i,d_j)$.

We now consider the **KD45-P** system including the above valid principles, *Modus ponens*(MP), as well as *Generalization* for the operator B.

3.4.8. DEFINITION. A model of **KD45-P** is a tuple $\langle W, D, R, \{\preceq_w\}_{w\in W}, V\rangle$, where W is a set of worlds, D is a set of constants, R is a euclidean and serial accessibility relation on W. Namely, it satisfies $\forall xyz((Rxy \wedge Rxz) \rightarrow Ryz)$ and $\forall x\exists y Rxy$. For each w, $\preceq_w$ is a quasi-linear order on D, which is the same throughout each euclidean class. V is evaluation function in an ordinary manner.

We remind the reader that in most respects euclidean classes are equivalence classes except that a number of points are irreflexive and have R relations just towards the reflexive members (the *equivalence part*) of the class.

3.4.9. THEOREM. *The* **KD45-P** *system is complete.*

Proof. The canonical model of this logic **KD45-P** has the required properties: The belief accessibility relation R is euclidean and serial. This means that with regard to R the model falls apart into euclidean classes. In each node $\underline{Pref}$ is a quasi-linear order of the constants. Within a euclidean class the preference order is constant (by $B\underline{Pref} \leftrightarrow \underline{Pref}$). This suffices to prove completeness. □

3.4.10. THEOREM. *The logic* **KD45-P** *has the finite model property.*

Proof. By standard methods. □

3.4.11. THEOREM. (*representation theorem*). $\vdash_{\mathbf{KD45-P}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences.*

Proof. Suppose that $\nvdash_{\mathbf{KD45-P}} \varphi(d_1, ..., d_n, p_1, ..., p_m)$. By Theorem 3.4.9, there is a model with a world w in which φ is falsified. We restrict the model to the euclidean class where w resides. Since the ordering of the constants is the same throughout euclidean classes, the ordering of the constants is now the same throughout the whole model. We can proceed as in Theorem 3.3.2 defining the predicates $P_1, \ldots, P_n$ in a constant manner throughout the model. □

3.4.12. REMARK. The three definitions above are not the only definitions that might be considered. For instance, we can give a variation (*) of Definition 3.4.2. For simplicity, we just use one predicate C.

$$Pref(x, y) ::= \neg B\neg C(x) \wedge B\neg C(y). \quad (*)$$

This means the agent can decide on her preference in a situation in which on the one hand she is not totally ready to believe $C(x)$, but considers it consistent with what she assumes, on the other hand, she distinctly believes $\neg C(y)$. Compared with Definition 3.4.2, (*) is weaker in the sense that it does not require explicit positive beliefs concerning $C(x)$.

We can even combine Definition 3.4.1 and (*), obtaining the following:

$$Pref(x, y) ::= (BC(x) \wedge \neg BC(x)) \vee (\neg B\neg C(x) \wedge B\neg C(y). \quad (**)$$

Contrary to (*), this gives a quasi-linear order.

Similarly, for Definition 3.4.3, if instead of $B(Supe(x, y))$, we use $\neg B\neg(Supe(x, y))$, a weaker preference definition is obtained.

3.5 Preference changes

So far we have given different definitions for preference in a stable situation. Now we direct ourselves to changes in this situation. In the definition of preference in the presence of complete information, the only item subject to change is the priority sequence. In the case of incomplete information, not only the priority sequence, but also our beliefs can change. Both changes in priority sequence and changes in belief can cause preference change. In this section we study both. Note that priority change leads to a preference change in a way similar to entrenchment change in belief revision theory (see [Rot03]), but we take the methodology of dynamic epistemic logic in this context.

3.5.1 Preference change due to priority change

Let us first look at a variation of Example 3.1.1:

3.5.1. Example. Alice won a lottery prize of ten million dollars. Her situation has changed dramatically. Now she considers the quality most important.

In other words, the ordering of the priorities has changed. We will focus on the priority changes, and the preference changes they cause. To this purpose, we start by making the priority sequence explicit in the preference. We do this first for the case of complete information in language without belief. Let $\mathfrak{C}$ be a priority sequence with length n as in Definition 3.2.1. Then we write $Pref_{\mathfrak{C}}(x, y)$ for the preference defined from that priority sequence. Let us consider the following possible changes: we write $\mathfrak{C}^\frown C$ for adding C to the right of $\mathfrak{C}$, $C^\frown\mathfrak{C}$ for adding C to the left of $\mathfrak{C}$, $\mathfrak{C}^-$ for the sequence $\mathfrak{C}$ with its final element deleted, and finally, $\mathfrak{C}^{i\leftrightarrows i+1}$ for the sequence $\mathfrak{C}$ with its i-th and $i{+}1$-th priorities switched. It is then clear that we have the following relationships:

$$
\begin{array}{l}
Pref_{\mathfrak{C}^\frown C}(x,y) \leftrightarrow Pref_{\mathfrak{C}}(x,y) \vee (Eq_{\mathfrak{C}}(x,y) \wedge C(x) \wedge \neg C(y)),\\
Pref_{C^\frown\mathfrak{C}}(x,y) \leftrightarrow (C(x) \wedge \neg C(y)) \vee ((C(x) \leftrightarrow C(y)) \wedge Pref_{\mathfrak{C}}(x,y)),\\
Pref_{\mathfrak{C}^-}(x,y) \leftrightarrow Pref_{\mathfrak{C},n-1}(x,y),\\
Pref_{\mathfrak{C}^{i\leftrightarrows i+1}}(x,y) \leftrightarrow Pref_{\mathfrak{C},i-1}(x,y) \vee (Eq_{\mathfrak{C},i-1}(x,y) \wedge C_{i+1}(x) \wedge \neg C_{i+1}(y)) \vee\\
(Eq_{\mathfrak{C},i-1}(x,y) \wedge (C_{i+1}(x) \leftrightarrow C_{i+1}(y)) \wedge C_i(x) \wedge \neg C_i(y)) \vee (Eq_{\mathfrak{C},i+1}(x,y) \wedge\\
Pref_{\mathfrak{C}}(x,y)).
\end{array}
$$

These relationships enable us to describe preference change due to changes of the priority sequence in the manner of dynamic epistemic logic. We now consider the following four operations: $[^+C]$ of adding C to the right, $[C^+]$ of adding C to the left, $[-]$ of dropping the last element of a priority sequence of length n, and $[i\leftrightarrow i{+}1]$ of interchanging the i-th and $i{+}1$-th elements. Then we obtain the following reduction axioms:

$[^{+}C]Pref(x,y) \leftrightarrow Pref(x,y) \vee (Eq(x,y) \wedge C(x) \wedge \neg C(y))$,
$[C^{+}]Pref(x,y) \leftrightarrow ((C(x) \wedge \neg C(y)) \vee ((C(x) \leftrightarrow C(y)) \wedge Pref(x,y)))$,
$[-]Pref(x,y) \leftrightarrow Pref_{n-1}(x,y)$,
$[i \leftrightarrow i+1]Pref(x,y) \leftrightarrow Pref_{i-1}(x,y) \vee (Eq_{i-1}(x,y) \wedge C_{i+1}(x) \wedge \neg C_{i+1}(y)) \vee (Pref_i(x,y) \wedge (C_{i+1}(x) \leftrightarrow C_{i+1}(y))) \vee (Eq_{i+1}(x,y) \wedge Pref(x,y))$.

Of course, the first two are the more satisfactory ones, as the right hand side is constructed solely on the basis of the previous $Pref$ and the added priority C. Note that one of the first two, plus the third and the fourth are sufficient to represent any change whatsoever in the priority sequence. Noteworthy also is that operator $[C^{+}]$ has exactly the same effects on a model as the operator $[\sharp C]$ in Chapter 2. We will discuss connections of this sort later in Chapter 4.

In the context of incomplete information when we have the language of belief, we can obtain similar reduction axioms for Definition 3.4.1 and 3.4.2. For instance, for Definition 3.4.1, we need only replace C by BC and $\neg C$ by $\neg BC$. For Definition 3.4.3, the situation is very complicated, reduction axioms are simply not possible. To see this, we return to the Example of Cora. Suppose Cora has a preference on the basis of cost and quality, and she also has the given information relating quality and neighborhood. Then her new preference after 'neighborhood' has been adjoined to the priority sequence is not a function of her previous preference and her beliefs about the neighborhood. The beliefs relating quality and neighborhood are central for her reasoning, but they are neither contained in the beliefs supporting her previous preference, nor in the beliefs about the neighborhood per se.

3.5.2 Preference change due to belief change

Now we move to the other source which causes preference change, namely, a change in belief. Such a thing often occurs in real life, new information comes in, one changes one's beliefs. Technically, the update mechanisms of [BS06a] and [Ben07a] can immediately be applied to our system with belief. As preference is defined in terms of beliefs, we can calculate preference changes from belief change. We distinguish the two cases that the belief change is caused by an update with so-called *hard* information and an update with *soft* information.

Preference change under hard information

Consider a simpler version of the Example 3.1.1:

3.5.2. Example. Let us assume that this time Alice only consider the houses' cost (C) and their neighborhood (N) with $C(x) \gg N(x)$. There are two houses d_1 and d_2 available. The real situation is that $C(d_1), N(d_1), C(d_2)$ and $\neg N(d_2)$. First Alice prefers d_2 over d_1 because she believes $C(d_2)$ and $N(d_1)$. However, now

Alice reads that $C(d_1)$ in a newspaper. She accepts this information. Accordingly, she changes her preference.

Here we assume that Alice treats the information obtained as hard information. She simply adds new information to her stock of beliefs. Figure 3.3 shows the situation before Alice's reading.

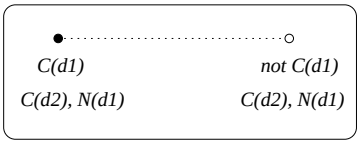

Figure 3.3: Initial model.

As usual, the dotted line denotes that Alice is uncertain about the two situations. In particular, she does not know whether $C(d_1)$ holds or not. After she reads that $C(d_1)$, the situation becomes Figure 3.4. The $\neg C(d_1)$-world is eliminated from the model: Alice has updated her beliefs. Now she prefers d_1 over d_2.

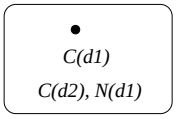

Figure 3.4: Updated model.

We have assumed that we are using the elimination semantics (e.g. [Ben06a], [FHMV95], etc.) in which public announcement of the sentence A leads to the elimination of the $\neg A$ worlds from the model. We have the reduction axiom:

$$[!A]Pref_{\mathfrak{C}}(x,y) \leftrightarrow A \rightarrow Pref_{A\rightarrow\mathfrak{C}}(x,y),$$

where, if $\mathfrak{C}$ is the priority sequence $C_1 \gg \cdots \gg C_n$, $A \rightarrow \mathfrak{C}$ is defined as $A \rightarrow C_1 \gg \cdots \gg A \rightarrow C_n$.

We can go even further if we use conditional beliefs $B^{\psi}\varphi$ as introduced in [Ben07a], with the meaning φ is believed under the condition of ψ. Naturally one can also introduce *conditional preference* $Pref^{\psi}(x,y)$, by replacing B in the definitions in Section 3.4 by B^{ψ}. Assuming A is a formula without belief operators, an easy calculation gives us another form of the reduction axiom:

$$[!A]Pref(x,y) \leftrightarrow A \rightarrow Pref^{A}(x,y).$$

Preference change under soft information

When incoming information is not as solid as considered in the above, we have to take into account the possibilities that the new information is not consistent with the beliefs the agent holds. Either the new information is unreliable, or the agent's beliefs are untenable. Let us switch to a semantical point of view for a moment. To discuss the impact of soft information on beliefs, the models are graded by a plausibility ordering $\leq$. For the one agent case one may just as well consider the model to consist of one euclidean class. The ordering of this euclidean class is such that the worlds in the equivalence part are the most plausible worlds. For all the worlds w in the equivalence part and all the worlds u outside it, $w < u$. Otherwise $v < v'$ can only obtain between worlds outside the equivalence part. To be able to refer to the elements in the model, instead of only to the worlds accessible by the R-relation, we introduce the universal modality U and its dual E. For the update by soft information, there are various approaches, we choose the *lexicographic upgrade* $\Uparrow A$ introduced by [Vel96] and [Rot06], adopted by [Ben07a] for this purpose. After the incoming information A, the ordering $\leq$ is updated by making all A-worlds strictly better than all $\neg A$-worlds keeping among the A-worlds the old orders intact and doing the same for the $\neg A$-worlds. After the update the R-relations just point to the best A-worlds. The reduction axiom for belief proposed in [Ben07a] is:

$$[\Uparrow A]B\varphi \leftrightarrow (EA \wedge B^A([\Uparrow A]\varphi) \vee (\neg EA \wedge B[\Uparrow A]\varphi)$$

We apply this only to priority formulas φ which do not have belief operators, and obtain for this restricted case a simpler form:

$$[\Uparrow A]B\varphi \leftrightarrow (EA \wedge B^A\varphi) \vee (\neg EA \wedge B\varphi).$$

From this one easily derive the reduction axiom for preference:

$$[\Uparrow A]Pref(x,y) \leftrightarrow (EA \wedge Pref^A(x,y)) \vee (\neg EA \wedge Pref(x,y)).$$

Or in a form closer to the one for hard information:

$$[\Uparrow A]Pref(x,y) \leftrightarrow (EA \rightarrow Pref^A(x,y)) \wedge (\neg EA \rightarrow Pref(x,y)).$$

The reduction axiom for conditional preference is:

$$[\Uparrow A]Pref^{\psi}(x,y) \leftrightarrow (E(A\wedge\psi) \rightarrow Pref^{A\wedge\psi}(x,y)) \wedge (\neg E(A\wedge\psi) \rightarrow Pref^{\psi}(x,y)).$$

By the fact that we have reduction axioms here, the completeness result in [Ben07a] for dynamic belief logic can be extended to a dynamic preference logic.

We will not spell out the details here.

3.6 Extension to the many agent case

This section extends the results of Section 3.4 to the many agent case. This will generally turn out to be more or less a routine matter. But at the end of the section, we will see that the priority base approach gives us a start of an analysis of cooperation and competition of agents. We consider agents here as cooperative if they have the same goals (priorities), competitive if they have opposite goals. This foreshadows the direction one may take to apply our approach to games. The language we are using is defined as follows.

3.6.1. Definition. Let Γ be a set of propositional variables, G be a group of agents, and D be a finite domain of objects, the *reduced language* of preference logic for many agents is defined in the following,

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \underline{Pref}^a(d_i, d_j) \mid B^a\varphi$$

where p, a, d_i respectively denote elements from Γ, G, and D.

Similarly to $\underline{Pref}^a$ expressing non-strict preference, we will use $Pref^a$ to denote the strict version. When we want to use the extended language, we add variables and the statements $P(d_i)$.

3.6.2. Definition. A *priority sequence* for an agent a is a finite ordered sequence of formulas written as follows: $C_1 \gg_a C_2 \cdots \gg_a C_n \quad (n \in \mathbb{N})$, where each C_m $(1 \leq m \leq n)$ is a formula from the language of Definition 3.6.1, with one single free variable x, but without $\underline{Pref}$ and B.

Here we take decisive preference to define an agent's preference. But the results of this section apply to other definitions just as well. It seems quite reasonable to allow in this definition of $Pref^a$ formulas that contain B^b and $Pref^b$ for agents b other than a. But we leave this for a future occasion.

3.6.3. Definition. Given a priority sequence of length n, two objects x and y, $Pref^a(x, y)$ is defined as follows:

$Pref^a_1(x, y) ::= B^aC_1(x) \wedge \neg B^aC_1(y)$,
$Pref^a_{k+1}(x, y) ::= Pref^a_k(x, y) \vee (Eq_k(x, y) \wedge B^aC_{k+1}(x) \wedge \neg B^aC_{k+1}(y)), k < n$,
$Pref^a(x, y) ::= Pref^a_n(x, y)$,

where $Eq_k(x, y)$ stands for $(B^aC_1(x) \leftrightarrow B^aC_1(y)) \wedge \cdots \wedge (B^aC_k(x) \leftrightarrow B^aC_k(y))$.

3.6.4. DEFINITION. The preference logic for many agents **KD45-P$^{\mathbf{G}}$** is consists of the following principles,

(a) $\underline{Pref^a}(d_i, d_i)$,
(b) $\underline{Pref^a}(d_i, d_j) \vee \underline{Pref^a}(d_j, d_i)$,
(c) $\underline{Pref^a}(d_i, d_j) \wedge \underline{Pref^a}(d_j, d_k) \rightarrow \underline{Pref^a}(d_i, d_k)$,
(1.) $\neg B^a \bot$,
(2.) $B^a\varphi \rightarrow B^a B^a \varphi$,
(3.) $\neg B^a\varphi \rightarrow B^a \neg B^a \varphi$,
(4.) $\underline{Pref^a}(d_i, d_j) \leftrightarrow B^a \underline{Pref^a}(d_i, d_j)$.

As usual, it also includes *Modus ponens*(MP), as well as *Generalization* for the operator B^a. It is easy to see that the above principles are valid for $\underline{Pref^a}$ extracted from a priority sequence.

3.6.5. THEOREM. *The preference logic for many agents* **KD45-P$^{\mathbf{G}}$** *is complete.*

Proof. The canonical model of this logic **KD45-P$^{\mathbf{G}}$** has the required properties: The belief accessibility relation R_a is euclidean and serial. This means that with regard to R_a the model falls apart into a-euclidean classes. Again, in each node $Pref^a$ is a quasi-linear order of the constants and within an a-euclidean class the a-preference order is constant. This quasi-linearity and constancy are of course the required properties for the preference relation. Same for the other agents. This shows completeness of the logic. □

3.6.6. THEOREM. *The logic* **KD45-P$^{\mathbf{G}}$** *has the finite model property.*

Proof. By standard methods. □

Similarly, a representation theorem can be obtained by showing that the model could have been obtained from priority sequences $C_1 \gg_a C_2 \cdots \gg_a C_m (m \in \mathbb{N})$ for all the agents.

3.6.7. THEOREM. (*representation theorem*). $\vdash_{\mathbf{KD45-P^G}} \varphi$ *iff* φ *is valid in all models with each* $\underline{Pref^a}$ *obtained from a priority sequence.*

Proof. Let there be k agents $a_0, \ldots, a_{k-1}$ and suppose $\varphi(d_1, \ldots, d_n)$. We provide each agent a_j with her own priority sequence $P_{n\times j+1} \gg_{a_j} P_{n\times j+2} \gg_{a_j} \ldots \gg_{a_j} P_{n\times(j+1)}$. It is sufficient to show that any model for **KD45-P$^{\mathbf{G}}$** for the reduced language can be extended by valuations for the $P_j(d_i)$'s in such a way that the preference relations are preserved. For each a_i-euclidean class, we follow the same procedure for $d_1, \ldots, d_n$ w.r.t. $P_{n\times j+1}, P_{n\times j+2}, \ldots, P_{n\times(j+1)}$ as in Theorem 3.3.2 w.r.t $P_1, \ldots, P_n$. The preference orders obtained in this manner are exactly the $Pref^{a_j}$ relations in the model. □

In the above case, the priority sequences for different agents are separate, and thus very different. Still stronger representation theorems can be obtained by requiring that the priority sequences for different agents are related, e.g. in the case of *cooperative agents* that they are equal. We will consider the two agent case in the following.

3.6.8. THEOREM. (*for two cooperative agents*). $\vdash_{\mathbf{KD45-P^G}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences shared by two cooperative agents.*

Proof. The 2 agents are a and b. We now have the priority sequence $P_1 \gg_a P_2 \gg_a ... \gg_a P_n$, same for b. It is sufficient to show that any model $\mathcal{M}$ with worlds W for **KD45-P$^{\mathbf{G}}$** for the reduced language can be extended by valuations for the $P_j(d_i)$'s in such a way that the preference relations are preserved. We start by making all $P_j(d_i)$'s true everywhere in the model. Next we extend the model as follows. For each a-euclidean class E in the model carry out the following procedure. Extend $\mathcal{M}$ with a complete copy $\mathcal{M}_E$ of $\mathcal{M}$ for all of the reduced language i.e. without the predicates P_j. Add R_a relations from any of the w in E to the copies v_E such that $w\, R_a\, v$. Now carry out the same procedure as in the proof of Theorem 3.3.2 in E's copy E_E. What we do in the rest of $\mathcal{M}_E$ is irrelevant. Now, in w, a will believe in $P_j(d_i)$ exactly as in the model in the previous proof, the overall truth of $P_j(d_i)$ in the a-euclidean class E in the original model has been made irrelevant. The preference orders obtained in this manner are exactly the $Pref^a$ relations in the model. All formulas in the reduced language keep their original valuation because the model $\mathcal{M}_E$ is bisimilar for the reduced language to the old model $\mathcal{M}$ as is the union of $\mathcal{M}$ and $\mathcal{M}_E$.

Finally do the same thing for b: add for each b-euclidean class in $\mathcal{M}$ a whole new copy, and repeat the procedure followed for a. Both a and b will have preferences with regard to the same priority sequence. □

For *competitive agents* we assume that if agent a has a priority sequence $D_1 \gg_a D_2 \gg \cdots \gg_a D_m (m \in \mathbb{N})$, then the opponent b has priority sequence $\neg D_m \gg_b \neg D_{m-1} \gg \cdots \gg_b \neg D_1$.

3.6.9. THEOREM. (*for two competitive agents*). $\vdash_{\mathbf{KD45-P^G}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences for competitive agents.*

Proof. Let's assume two agents a and b. For a we take a priority sequence $P_1 \gg_a P_2 \gg_a \cdots \gg_a P_n \gg_a P_{n+1} \gg_a \cdots \gg_a P_{2n}$, and for b, we take $\neg P_{2n} \gg_b \neg P_{2n-1} \gg_b \cdots \gg_b \neg P_n \gg_b \neg P_{n-1} \gg_b \cdots \gg_b \neg P_1$. It is sufficient to show that any model $\mathcal{M}$ with worlds W for **KD45-P$^{\mathbf{G}}$** for the reduced language can be extended by valuations for the $P_j(d_i)$'s in such a way that the preference relations are preserved. We start by making all $P_1(d_i) \ldots P_n(d_i)$ true everywhere in the model and $P_{n+1}(d_i) \ldots P_{2n}(d_i)$ all false everywhere in the model. Next we extend the model as follows.

For each a-euclidean class E in the model carry out the following procedure. Extend $\mathcal{M}$ with a complete copy $\mathcal{M}_E$ of $\mathcal{M}$ for all of the reduced language i.e. without the predicates P_j. Add R_a relations from any of the w in E to the copies v_E such that $w\,R_a\,v$. Now define the values of the $P_1(d_i)\ldots P_n(d_i)$ in E_E as in the previous proof and make all $P_m(d_i)$ true everywhere for $m > n$. The preference orders obtained in this manner are exactly the $Pref^a$ relations in the model.

For each b-euclidean class E in the model carry out the following procedure. Extend $\mathcal{M}$ with a complete copy $\mathcal{M}_E$ of $\mathcal{M}$ for all of the reduced language i.e. without the predicates P_j. Add R_b relations from any of the w in E to the copies v_E such that $w\,R_b\,v$. Now define the values of the $\neg P_{2n}(d_i)\ldots\neg P_{n+1}(d_i)$ in E_E as for $P_1(d_i)\ldots P_n(d_i)$ in the previous proof and make all $P_m(d_i)$ true everywhere for $m \leq n$. The preference orders obtained in this manner are exactly the $Pref^b$ relations in the model.

All formulas in the reduced language keep their original valuation because the model $\mathcal{M}_E$ is bisimilar for the reduced language to the old model $\mathcal{M}$ as is the union of $\mathcal{M}$ and all the $\mathcal{M}_E$. □

3.6.10. Remark. These last representation theorems show that they are as is to be expected not only a strength but also a weakness. The weakness here is that they show that cooperation and competition cannot be differentiated in this language. On the other hand, the theorems are not trivial, one might think for example that if a and b cooperate, $B_aPref_b(c,d)$ would imply $Pref_a(c,d)$. This is of course completely false, a and b can even when they have the same priorities have quite different beliefs about how the priorities apply to the constants. But the theorems show that no principles can be found that are valid only for cooperating agents. Moreover they show that if one wants to prove that $B_aPref_b(c,d) \rightarrow Pref_a(c,d)$ is not valid for cooperating agents a counterexample to it in which the agents do not cooperate suffices.

3.7 Preference over propositions

Most other authors on preference have discussed preference over propositions rather than objects. Our approach can be applied to preference over propositions as well. We are going to develop this ideas further in this section. As we know, preference is always intertwined with beliefs. In the following, we will propose a system combining them. And we specially take the line that preference is a state of mind and that therefore one prefers one alternative over another if and only if one believes one does. If we take this line, the most obvious way would be to go to second order logic and consider priority sequence $A_1(\varphi) \gg A_2(\varphi) \gg \ldots, \gg A_n(\varphi)$, where the A_i are properties of propositions. However, we find it close to our intuitions to stay first order as much as possible. With that in mind, we define the new priority sequence for the propositional case as follows.

3.7.1. Definition. A *propositional priority sequence* is a finite ordered sequence of formulas written as follows

$$\varphi_1(x) \gg \varphi_2(x) \gg \cdots \gg \varphi_n(x) \quad (n \in \mathbb{N})$$

where each of $\varphi_m(x)$ is a propositional formula with an additional propositional variable, x, which is a common one to each $\varphi_m(x)$.

Formulas $\varphi(x)$ can express properties of propositions, for instance, applied to ψ, $x \rightarrow p_1$ expresses that ψ implies p_1, "ψ has the property" p_1.

We apply our approach in previous sections to define preference in terms of beliefs. As we have seen in Section 3.4, there are various ways to do it. We are guided by the definition of decisive preference in formulating the following:

3.7.2. Definition. Given a propositional priority sequence of length n, we define preference over propositions ψ and θ as follows:

$$Pref(\psi, \theta) \quad \text{iff} \quad \text{for some } i \ (B(\varphi_1(\psi) \leftrightarrow B(\varphi_1(\theta)) \wedge \cdots \wedge (B(\varphi_{i-1}(\psi)) \leftrightarrow B(\varphi_{i-1}(\theta))) \wedge (B(\varphi_i(\psi) \wedge \neg B(\varphi_i(\theta))$$

Note that preference between propositions is in this case almost a preference between mutually exclusive alternatives: in the general case one can conclude beyond the quasi-linear order that derives directly from our method only that if $B(\psi \leftrightarrow \theta)$, then ψ and θ are equally preferable. Otherwise, any proposition can be preferable over any other.

For some purposes (this will get clearer in the proof of the representation theorem below), we need a further generalization, hence here we give a slightly more complex definition.

3.7.3. Definition. A *propositional priority sequence* is a finite ordered sequence of sets of formulas written as follows

$$\Phi_1 \gg \Phi_2 \gg \cdots \gg \Phi_n$$

where each set Φ_i consists of propositional formulas that have an additional propositional variable, x, which is a common one to each Φ_i.

A new definition of preference is given by:

3.7.4. Definition. Given a propositional priority sequence of length n, we define preference over propositions ψ and θ as follows:

$$Pref(\psi, \theta) \quad \text{iff} \quad \exists i(\forall j < i(\exists \varphi \in \Phi_j(B\varphi(\psi)) \leftrightarrow \exists \varphi \in \Phi_j(B\varphi(\theta) \wedge \exists \varphi \in \Phi_i(B\varphi(\psi)) \wedge \forall \varphi \in \Phi_i \neg B(\varphi\theta)))$$

3.7.5. Remark. In fact, the priority set Φ_m could be expressed by one formula

$$\bigvee_{\varphi\in\Phi_m} B\varphi.$$

But then we would have to use B in the formulas of the priority sequence, which we prefer not to.

The axiom system **BP** that arises from these considerations combines preference and beliefs in the following manner:

(a) $\underline{Pref}(\varphi,\varphi)$

(b) $\underline{Pref}(\varphi,\psi)\wedge\underline{Pref}(\psi,\theta)\rightarrow\underline{Pref}(\varphi,\theta)$

(c) $\underline{Pref}(\varphi,\psi)\vee\underline{Pref}(\psi,\varphi)$

(d) $B\underline{Pref}(\varphi,\psi)\leftrightarrow\underline{Pref}(\varphi,\psi)$

(e) $B(\varphi\leftrightarrow\psi)\rightarrow\underline{Pref}(\varphi,\psi)\wedge\underline{Pref}(\psi,\varphi)$.

As usual, it also includes *Modus ponens (MP)*, as well as the Generalization Rule for the operator B. The first three are standard for preference, and we have seen the analogue of (d) in Section 3.4. (e) is new, as a connection between beliefs and preference. It expresses that if two propositions are indistinguishable on the plausible worlds they should be equally preferable. It is easy to see that the above axioms are valid in the models defined as follows.

3.7.6. Definition. A model of **BP** is a tuple $\langle W,R,\{\preceq_w\}_{w\in W},V\rangle$, where W is a set of worlds, R is a euclidean and serial accessibility relation on W. Namely, it satisfies $\forall xyz((Rxy\wedge Rxz)\rightarrow Ryz)$ and $\forall x\exists yRxy$. Moreover, for each w, $\preceq_w$ is a quasi-linear order on propositions (subsets of W), which is constant throughout each euclidean class and which is determined by the part of the propositions that lies within the 'plausibility part' of the euclidean class. V is an evaluation function in an ordinary manner.

3.7.7. Theorem. *The* **BP** *system is complete w.r.t the above models.*

Proof. Assume $\nvdash_{\mathbf{BP}}\theta$. Take the canonical model $\mathcal{M}=(W,R,V)$ for the formulas using only the propositional variables of θ. To each world of W a quasi-linear order of all formulas is associated, and it only depends on the extension of the formula (the set of nodes where the formula is true) in the plausible part of the model. This order is constant throughout the euclidean class defined by R. $\neg\theta$ can be extended to a maximal consistent set Γ. We consider the submodel generated by Γ, $\mathcal{M}'=(W',R,V)$, which naturally is an euclidean class. Since each world in W' has access to the same worlds, each world that satisfies the same atoms

satisfies the same formulas. In fact, each formula φ in this model is equivalent to a purely propositional formula, a formula without B or $Pref$. To see this, one just has to realize that $B\psi$ is in the model either equivalent to $\top$ or $\bot$, and the same holds for $Pref(\psi, \theta)$. (Note that this argument only applies because we have just one euclidean class.) Now apply a p-morphism to $\mathcal{M}'$ which identifies worlds that satisfy the same formula. This gives a finite model consisting of one euclidean class with a constant order that still falsifies θ. Moreover, each world is characterized by a formula $\pm p_1 \wedge \cdots \wedge \pm p_k$ that expresses which atoms are true in it. In consequence, each subset of the model (proposition) is also definable by a purely propositional formula, a disjunction of the formulas $\pm p_1, \wedge \cdots \wedge \pm p_k$ describing its elements. □

Similarly, we can prove the representation result.

3.7.8. Theorem. (*representation theorem*) $\nvdash_{\mathbf{BP}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences.*

Proof. The order of the finitely many formulas defining all the subsets of the models can be represented as a sequence

$$\Phi_1, \ldots, \Phi_k$$

where Φ_1 are the best propositions ($\varphi, \psi \in \Phi_1$ implies $\varphi \trianglelefteq \psi$ and $\psi \trianglelefteq \varphi$, Φ_i are the next best propositions, etc. Then the following is the priority sequence which results in the given order:

$$\{x \leftrightarrow \varphi \mid \varphi \in \Phi_1\} \gg \cdots \gg \{x \leftrightarrow \varphi \mid \varphi \in \Phi_k\}.$$

□

So far our discussions on the preference relation over propositions are rather general. We do not presuppose any restriction on such a relation. However, if we think that the preference relation over propositions is a result of lifting a preference relation over possible worlds (as discussed before), we specify its meaning in a more precise way, following the obvious option of choosing different combinations of quantifiers. For example, we can take $\forall\exists$ preference relations over the propositions, i.e. preference relations over propositions lifted from preference relations over worlds in the $\forall\exists$ manner. Regarding the axiomatization, we will then have to add the following two axioms to the above **BP** system, the new system will denoted as $\mathbf{BP}^{\forall\exists}$. It has two more axioms:

- $B(\varphi \to \psi) \to \underline{Pref}(\psi, \varphi)$.
- $\underline{Pref}(\varphi, \varphi_1) \wedge \underline{Pref}(\varphi, \varphi_2) \to \underline{Pref}(\varphi, \varphi_1 \vee \varphi_2)$

3.7.9. THEOREM. *$BP^{\forall\exists}$ is complete.*

Proof. By an adaption of the proof by [Hal97]. The difference is: [Hal97] uses a combination of preference and universal modality. Instead, our system is a combination of belief and preference. This means what is preferred in our system is decided by the plausible part of the model. However, this will not affect the completeness proof much. □

3.7.10. REMARK. In fact, $[pref]\varphi$ in Chapter 2 can be defined now as $\underline{Pref}(\varphi, \top)$. Then the preference used in the system $\mathbf{BP}^{\forall\exists}$ is simply the following:

$$\underline{Pref}(\varphi, \psi) \leftrightarrow B(\psi \rightarrow \langle pref\rangle\varphi)$$

We will come back to this point in Chapter 4.

Similarly, we get the representation result for this restricted case:

3.7.11. THEOREM. (*representation theorem*) $\vdash_{\mathbf{BP}^{\forall\exists}} \varphi$ *iff* φ *is valid in all* $\forall\exists$*-models obtained from priority sequences.*

The proof is same as for the basic system.

Finally, to conclude this subsection, recall that we had a logic system to discuss preference over objects when beliefs are involved. With our new system just presented, we can talk about preference over propositions. But what is the relation between these two systems? The following theorem provides an answer.

3.7.12. THEOREM. $\vdash_{\mathbf{KD45-P}} \varphi(d_1, \ldots, d_n)$ *iff* $\vdash_{\mathbf{BP}} \varphi(p_1, \ldots, p_n)$ *where the propositional variables* $p_1, \ldots, p_n$ *do not occur in* $\varphi(d_1, \ldots, d_n)$.

Proof. In order to prove this theorem, we need to prove the following lemma:

3.7.13. LEMMA. *If* $\nvdash_{\mathbf{KD45-P}} \varphi(d_1, \ldots, d_n)$, *then for each* n *there is a model* $\mathcal{M} \models \neg\varphi$ *with at least* n *elements.*

Proof. Assume that we only have a model $\mathcal{M} = (W, R, V)$ in which W has m elements, where $m < n$. Take one element of W, say w, and make copies of it, say, w_1, w_2,..., w_k, till we get at least n elements. If wRv, then we make w_iRv, and if vRw, then vRw_i. In this way we get a new model with at least n elements. It is bisimilar to the original model. □

Now we are ready to prove the theorem.
($\Rightarrow$). It is easy to see that all the **KD45-P** axioms and rules are valid in **BP** if one replaces each d_i by p_i.

($\Leftarrow$). It is sufficient to transform any finite **KD45-P** model $\mathcal{M}$ with only one euclidean class into a **BP** model $\mathcal{M}'$ with at least n possible worlds in which for each w and each ψ, $\mathcal{M}', w \models \psi(p_1, \ldots, p_n)$ iff $\mathcal{M}, w \models \psi(d_1, \ldots, d_n)$. Let $\mathcal{M} = (W, R, \preceq, V)$, then $\mathcal{M}' = (W', R, \trianglelefteq, V')$, where V' is like V except that for the $p_1, \ldots, p_n$, we assign $V'(p_i) = V'(p_j)$ if $d_i \preceq d_j \wedge d_j \preceq d_i$, otherwise, $V'(p_i) \neq V'(p_j)$. [5] According to Lemma 3.7.13, there are enough subsets to do this. Finally, we set $V'(p_i) \lhd V'(p_j)$ iff $d_i \prec d_j$ and extend $\lhd$ to other sets in an arbitrary manner. □

If one thinks of propositional variables as representing basic propositions, then this theorem says that reasoning about preference over objects is the same as reasoning about preference over basic propositions. This is not surprising if one thinks of basic propositions as exclusive alternatives as are objects. Of course, the logic of preference over propositions in general is more expressive. One can look at this latter fact in two different ways: (a) one may think the logic over preference over all propositions as essentially richer than the logic of the basic propositions or objects, or (b) one may think that the essence of the logic of propositions is contained in the basic propositions (represented by the propositional variables) and the rest needs to be carried along in the theory to obtain a good logical system but is of little value by itself.

By applying the method of [Hal97] we can adapt the above proof to obtain the following:

3.7.14. THEOREM. $\vdash_{\mathbf{KD45-P}} \varphi(d_1, \ldots, d_n)$ *iff* $\vdash_{\mathbf{BP}^{\forall\exists}} \varphi(p_1, \ldots, p_n)$ *where the propositional variables* $p_1, \ldots, p_n$ *do not occur in* $\varphi(d_1, \ldots, d_n)$.

Up to now we have used decisive preference. Another option is to use deliberate preference. Let us look at this in a rather general manner. Assume that $Supe(\varphi, \psi)$ has the property in a model that for each φ, ψ,

$$\models (\varphi \leftrightarrow \varphi') \wedge (\psi \leftrightarrow \psi') \rightarrow (Supe(\varphi, \psi) \leftrightarrow Supe(\varphi', \psi')),$$

we then say 'superior' is a *local property* in that model. We can now state the following propositions.

3.7.15. THEOREM. *If we define* $Pref(\varphi, \psi)$ *as* $B(Supe(\varphi, \psi))$ *in any model where* $Supe(\varphi, \psi)$ *is a local partial order, then* $Pref(\varphi, \psi)$ *satisfies the principles of* **BP**, *except possibly connectedness.*

It is to be noted that

$$\varphi \rightarrow \langle pref \rangle \psi$$

[5] Note that the $V'(p_i)$ are only relevant for the ordering $\trianglelefteq$ because the p_i's only occur directly under the $Pref$ in $\varphi(p_1, \ldots, p_n)$.

is not a local property even if $\preceq$ is a subrelation of R. Nevertheless, in case $\preceq$ is a subrelation of R, $B(\varphi \to \langle pref \rangle \psi)$ does satisfy the principles of **BP** minus connectedness, and the additional $\mathbf{BP}^{\forall\exists}$ axioms, as we commented in Remark 3.7.10. For this purpose the following weakening of locality is sufficient:

$$\models (\varphi \leftrightarrow \varphi') \wedge B(\varphi \leftrightarrow \varphi') \wedge (\psi \leftrightarrow \psi') \wedge B(\psi \leftrightarrow \psi') \to (Supe(\varphi, \psi) \leftrightarrow Supe(\varphi', \psi')).$$

3.8 Discussion and conclusion

Partially ordered priority sequence

A new situation occurs when there are several priorities of incomparable strength. Take the Example 3.1.1 again, however, instead of considering three properties, Alice also takes the 'transportation convenience' into her account. But for her neighborhood and transportation convenience are really incomparable. Abstractly speaking, this means that the priority sequence is now *partially ordered*. We show in the following how to define preference based on a partially ordered priority sequence. In other words, we consider a set of priorities $C_1 ..., C_n$ with the relation $\gg$ between them a partial order.

3.8.1. Definition. We define $Pref_n(x, y)$ by induction, where $\{n_1, ..., n_k\}$ is the set of immediate predecessors of n.

$$\underline{Pref_n(x,y)} ::= \underline{Pref_{n_1}}(x,y) \wedge ... \wedge \underline{Pref_{n_k}}(x,y) \wedge ((C_n(y) \to C_n(x)) \vee (Pref_{n_1}(x,y) \vee ... \vee Pref_{n_k}(x,y)))$$

where as always $Pref_m(x, y) \leftrightarrow \underline{Pref_m}(x, y) \wedge \neg \underline{Pref_m}(y, x)$

This definition is, for finite partial orders, equivalent to the one in [Gro91] and [ARS02]. More discussion on the relation between partially ordered priorities and G-spheres, see [Lew81]. When the set of priorities is unordered, again, we refer to [Kra81]. We come back to this issue in Chapter 4.

Conclusion

In this chapter we considered preference over objects. We showed how this preference can be derived from priorities, properties of these objects. We did this both in the case when an agent has complete information and in the case when an agent only has beliefs about properties. We considered both the single and multi-agent case. In all cases, we constructed preference logics, some of them extending the standard logic of belief. This leads to interesting connections between preference and beliefs. We strengthened the usual completeness results for logics of this kind to representation theorems. The representation theorems describe the reasoning

that is valid for preference relations that have been obtained from priorities. In the multi-agent case, these representation theorems are strengthened to special cases of cooperative and competitive agents. We studied preference change with regard to changes of the priority sequence, and change of beliefs. We applied the dynamic epistemic logic approach, and in consequence reduction axioms were presented. We proposed a new system combining preference and beliefs, talking about preference over propositions. We concluded by some discussion on generalizing the linear orders in this chapter to partial orders.

Chapter 4

Comparisons and Combinations

In the preceding two chapters, we have presented two different approaches to preference structure and preference change. These proposals were based on different intuitions, both plausible and attractive. Even so, the question naturally arises how the two perspectives are related. The aim of the present chapter is to draw a comparison, connect the modal logic based view of Chapter 2 with the priority-based view of Chapter 3, and try to integrate them. To see why this makes sense, let us start by briefly summarizing some key ideas.

The approach taken in Chapter 2 had the following main points:

- The *basic structures* were models $(W, \sim, \preceq, V)$ with a set of worlds (or objects) with a reflexive and transitive binary 'betterness' relation $\preceq$ ('at least as good as'), while we also assumed a standard epistemic accessibility relation for agents.

- The *language* used was an epistemic language extended with a universal modality U (or its existential dual E) plus a standard unary modality $[bett]$ using betterness as its accessibility relation.[1]

- *Preference* was treated as a relation over propositions, with the latter viewed as sets of possible worlds. Precisely, it is a *lifting* of the betterness relation to such sets. There are various ways of lifting, determined by quantifier combinations. One typical example uses $\forall\exists$, and it was defined as:

$$Pref^{\forall\exists}(\varphi, \psi) ::= U(\psi \rightarrow \langle bett \rangle \varphi).$$

  An alternative would be to take the epistemic modality K instead of U here, making preference a partly betterness-based, partly epistemic notion, subject to introspection in the usual way.

[1] To distinguish from notations we will use later on for preference, we write the operator here as $[bett]$ instead of $[pref]$ in Chapter 2. We will mostly omit the subscript for agents.

- *T*he language also had a *dynamic* aspect, and information update affects knowledge just as in standard DEL. Likewise, changes in preference were dealt with, by first defining changes in the basic betterness relations at the possible world level. These were handled by the standard DEL methodology. A typical example was the new reduction axiom for changes in betterness modalities after the action of 'suggesting that A':

$$\langle\sharp\varphi\rangle\langle bett\rangle\psi \leftrightarrow (\neg\varphi \wedge \langle bett\rangle\langle\sharp\varphi\rangle\psi) \vee (\langle bett\rangle(\varphi \wedge \langle\sharp\varphi\rangle\psi)).$$

  Given this reduction axiom plus that for the universal modality U, we were then able to derive a reduction axiom for the propositional preference operator $Pref^{\forall\exists}$. In case we use the epistemic modality to define preference between propositions, this will actually make *two* types of event, and two corresponding reduction axioms relevant. In addition to explicit betterness transformers such as suggestions, also, a pure information update affects K operators, and hence also the preferences involving them.

Thus, Chapter 2 provides an account of betterness ordering of objects, and its dynamics, intertwined with agents' knowledge and information update. Preference comes out as a *defined* concept.

Prima facie, Chapter 3 took a quite different approach, starting from a given priority order among propositions ('priorities'), and then deriving a preference order among objects. Again, we review the main contributions:

- A *priority base* is given first, consisting of strictly ordered properties:

$$P_1(x) \gg P_2(x) \gg \cdots \gg P_n(x)$$

  Next, a *preference order* $\preceq$ over objects is *derived* from this priority base, depending on whether the objects have the properties in the priority sequence or not. There are many ways for such a derivation, but we have taken one inspired by Optimality Theory (OT) saying that the earlier priorities in the given sequence count strictly heavier than the later ones. The preference derived this way is a quasi-linear order, not just reflexive and transitive, but also 'connected'.

- Here, too, preference was intertwined with information and agents' propositional attitudes, but this time, focused on their *beliefs*. In the case of incomplete information about which properties in the priority sequence objects possess, preference was defined in terms of which properties agents believe the objects to possess.

- On the syntax side, to speak about preference over objects, a fragment of a *first-order language* has been used, with a binary relation $Pref$. A doxastic belief operator was then added to explicitly describe object preferences based on beliefs.

- In this setting, too, we analyzed *changes in preference*. There are two possible sources for this, as before. Either preference change is caused by a change in the priority sequence, leading to a new way of ordering objects, or it is caused by a change in beliefs. We have given reduction axioms for both these scenarios, now for the languages appropriate here.

Clearly, despite the difference in starting point, the agendas of Chapters 2 and 3 are very similar. The purpose of this Chapter is to make this explicit, and see what questions arise when we make the analogies more precise.

Our discussion will be mostly semantics-oriented, though we will get to a comparison with syntax later in this chapter. Next, we will make the following simplification, or rather abstraction. At the surface, Chapter 2 speaks about ordering over possible worlds, and propositions as set of possible worlds, while Chapter 3 is about ordering 'individual objects' using their properties. In what follows, we will take all 'objects' to be 'worlds' in modal models - but this is just a vivid manner of speaking, and nothing would be lost if the reader were just to think of 'points' and 'properties' instead of worlds and propositions.[2] Finally, in order to be neutral on the different perspectives of Chapters 2 and 3, we start with two orderings at different levels. One is the betterness relation over possible worlds, written as $(W, \preceq)$, the other a preference or priority relation over propositions, viewed as sets of possible worlds, denoted by $(\mathscr{P}, <)$.

The greater part of this chapter will be devoted to exploring the deeper connection between these two orderings. The main questions we are going to pursue are the following:

- How to *derive* a preference order of 'betterness' over possible worlds from an ordered priority sequence?

- In the opposite direction, how to *lift* a betterness relation on worlds to an ordering over propositions?

The following diagram shows these two complementary directions:

$$\begin{array}{ccc} & (\mathscr{P}, <) & \\ \text{lift} \uparrow & & \downarrow \text{derive} \\ & (W, \preceq) & \end{array}$$

[2]There are interesting intuitive differences, however, between *object preference* and *world preference*, which will be discussed briefly at the end of this chapter. See also the remark after Theorem 3.7.12.

Besides these connections between the two levels, what will be of interest to us is how to relate dynamical changes at the two levels to each other.

The chapter is organized as follows. In Section 4.1 we propose a simple structure called *structured model* containing both a preference over possible worlds, and an ordering over propositions. In Section 4.2, we first study ways of deriving object preferences from a priority base, including some representation theorems. We will relate our method to other approaches in the literature, in particular, the preference merge in [ARS02]. In Section 4.3, we look at the opposite direction: how to lift an object preference relation to an ordering over propositions. A characterization theorem will be proved for the natural lifting of type $\forall\exists$. In Section 4.4 we study how concrete order-changing operations at the two levels correspond to each other. Section 4.5 is to connect the PDL-definable preference change to an alternative approach from the recent literature, product update on belief revision, which uses 'event model' as in dynamic-epistemic logic. Then in Section 4.6, we move from semantic structures to formal languages, and provide a comparison of the various logical languages used in the previous two chapters. Finally, in Section 4.7 we compare the different ways of preference interacting with belief in the previous two chapters, and we end this chapter with a proposal of putting all our systems together in one 'doxastic preferential predicate logic' of object and world preference.

4.1 Structured models

Preference over propositions $(\mathscr{P}, <)$ and betterness over possible worlds $(W, \preceq)$ can be brought together as follows:

4.1.1. Definition. A *structured model* $\mathcal{M}$ is a tuple $(W, \preceq, V, (\mathscr{P}, <))$, where W is a set of possible worlds, $\preceq$ a preference relation over W, V a valuation function for proposition letters, and $(\mathscr{P}, <)$ an ordered set of propositions, the 'important properties' or priorities.[3]

Structured models extend standard modal models, which may be viewed as the special case where $\mathscr{P}$ equals the powerset of W.

Here are some further notational stipulations. As in Chapter 2, $y \preceq x$ means that the world x is 'at least as good as' the world y or 'preferable over' y, while

[3]Compared with ordinary modal models $(W, \preceq, V)$, structured models have a new component, viz. a set of distinguished propositions $\mathscr{P}$. Agenda-based modal models introduced in [Gir08] have a similar structure, as an agenda is a set of distinguished propositions, too. It would then be very natural to look at modal languages over worlds where the valuation map only assigns propositions from $\mathscr{P}$ as values to atomic proposition letters. Moreover, we can let $\mathscr{P}$ determine the world preference relations. It would be interesting to find out what happens to standard modal logics on such restricted models, as these will now encode information about the structure of $\mathscr{P}$. This issue will not be pursued in this chapter.

$y \prec x$ means that x is strictly preferable over y: i.e. $y \preceq x$ but not $x \preceq y$. To emphasize preference relations induced by a priority sequence $(\mathscr{P}, <)$, we will write $y \preceq_{\mathscr{P}} x$. In general, the set $\mathscr{P}$ will be a partial order - but it is useful to also have a simpler case as a warm-up example. Suppose that $\mathscr{P}$ is a flat set, without any ordering. We then write the structured model simply as $(W, \preceq, V, \mathscr{P})$.

Some explanations on notation: We use capital letters $S, T, X, Y, P_i, \ldots$ for arbitrary propositions in the set $\mathscr{P}$, small letters $x, y, z, \ldots$ for arbitrary possible worlds, while Px means that $x \in P$. As for the other level, we write $Y \trianglelefteq X$ (or $Y \triangleleft X$) when proposition X is at least as good as (or strictly preferable to) Y.

Structured models simply combine the approaches in the previous two chapters. In particular, the syntactic priorities of Chapter 3 are now moved directly into the models, and sit together with an order over worlds. We will see how these two layers are connected in the next sections.

4.2 Moving between levels: From propositional priorities to object preferences

We first look at the derivation of preference ordering from a primitive priority sequence. This is a common scenario in many research areas. For instance, given a goal base as a finite set of propositions with an associated rank function, [CMLLM04] extends this priority on goals to a preference relation on alternatives. Likewise, in the theory of belief revision, an epistemic 'entrenchment relation' orders beliefs, those with the lowest entrenchment being the ones that are most readily given up ([GM88], [Rot03]). But our main motivating example in Chapter 3 was linguistic *Optimality Theory* ([PS93]). Here a set of alternative structures is generated by the grammatical or phonological theory, while an order over that set is determined by given strictly ordered constraints. Language users then employ the optimal alternative that satisfies the relevant constraints best.

What interests us in this chapter is the formal mechanism itself, i.e. how to get a preference order from a priority sequence. There are many proposals to this effect in the literature. In what follows, we will discuss a few, including the one adopted in Chapter 3, to place things in a more general perspective, and facilitate comparison with Chapter 2.

Chapter 3 considered only finite linearly ordered priority sequences, and object order was derived via an *OT*-style lexicographic stipulation. In terms of structured models $(W, \preceq, V, (\mathscr{P}, <))$, $(\mathscr{P}, <)$ is a finite linear order. Now, we first spell out the *OT*-definition:[4]

[4]The formulation here is slightly, but inessentially different from the inductive version we had in Chapter 3.

$$y \preceq^{OT}_{\mathscr{P}} x ::= \forall P{\in}\mathscr{P}(Px{\leftrightarrow}Py) \vee \exists P'{\in}\mathscr{P}(\forall P{<}P'(Px{\leftrightarrow}Py) \wedge (P'x{\wedge}\neg P'y)).$$

To recall how this works, we repeat an earlier illustration from Chapter 3:

4.2.1. Example. Alice is going to buy a house. In doing so, she has several things to consider: the cost, the quality, and the neighborhood. She has the following priority sequence:

$$C(x) \gg Q(x) \gg N(x),$$

where $C(x)$, $Q(x)$ and $N(x)$ stand for 'x has low cost', 'x is of good quality' and 'x has a nice neighborhood', respectively. Consider two houses d_1 and d_2 with the following properties: $C(d_1), C(d_2), \neg Q(d_1), \neg Q(d_2), N(d_1)$ and $\neg N(d_2)$. According to the OT-definition, Alice prefers d_1 over d_2 strictly, i.e. $d_2 \prec d_1$.

This OT-definition is by no means new. It was also investigated in other literature, e.g. [BCD+93] on priority-based handling of inconsistent sets of classical formulas, and [Leh95] on getting new conclusions from a set of default propositions. It has been called *leximin ordering* in this and other literature.

But besides the OT-definition, there are other ways of deriving object preferences from a priority base. To see this, first consider the above-mentioned simplest case where $\mathscr{P}$ is *flat*. The following definition gives us a very natural order (we call it the '$*$-definition'):

$$y \preceq^{*}_{\mathscr{P}} x ::= \forall P \in \mathscr{P}(Py \rightarrow Px).$$

This is found, e.g. in the theory of default reasoning of Veltman ([Vel96]), or in the topological order theory of Chu Spaces ([Ben00]). Incidentally, the same order would arise on our OT-definition of object preference if we take the flat set to have the trivial universal ordering relation. Next, as noted earlier, given a non-strict order $\preceq$, one can define its strict version $\prec$ in the following:

$$y \prec x ::= y \preceq x \wedge \neg(x \preceq y).$$

So the strict version of $y \preceq^{*}_{\mathscr{P}} x$ can be written as:

$$y \prec^{*}_{\mathscr{P}} x ::= \forall P \in \mathscr{P}(Py \rightarrow Px) \wedge \exists P' \in \mathscr{P}(P'x \wedge \neg P'y).$$

In the following we will only present non-strict versions of preference.

Returning to general structured models with the OT-definition, several questions arise naturally. Which orders over possible worlds are produced? Can we always find some priority sequence that produces such a given object order? The following representation result gives a precise answer. We had a similar result in Chapter 3, but this time, we will provide a new proof, while dropping the finiteness assumption.

4.2.2. THEOREM. *For any standard model $\mathcal{M} = (W, \preceq, V)$, the following two statements are equivalent:*
(a) there is a structured model $\mathcal{M}' = (W, \preceq, V, (\mathscr{P}, <))$ s.t.

$$y \preceq x \quad \textit{iff} \quad y \preceq^{OT}_{\mathscr{P}} x \quad \textit{for all} \quad x, y \in W.$$

(b) $y \preceq x$ is a quasi-linear order.

Proof. (a)⇒(b). Chapter 3 showed that the OT-definition always generates a quasi-linear order $\preceq^{OT}$.
Now for the converse direction (b)⇒(a). First, define a 'cluster' as a maximal subset X of W such that $\forall y, z \in X$: $y \preceq z$. Clusters exist by Zorn's Lemma, and different clusters are disjoint by their maximality. Each point x belongs to a cluster, which we will call C_x. First, we define a natural ordering of clusters reflecting that of the worlds:

$$C' \trianglelefteq C \quad \text{if} \quad \exists y \in C', \exists x \in C : y \preceq x.$$

We first prove the following connection with the given underlying object order:

4.2.3. LEMMA.

$$y \preceq x \quad \textit{iff} \quad C_y \trianglelefteq C_x.$$

Proof. (⇒). By definition, $x \in C_x$ and $y \in C_y$, so $C_y \trianglelefteq C_x$.
(⇐). If $C_y \trianglelefteq C_x$, then by definition $\exists u \in C_x$, $v \in C_y$ with $v \preceq u$. So we have $u \preceq x (x \in C_x)$ and $y \preceq v (y \in C_y)$ – and hence by transitivity, we get $y \preceq x$. □

Importantly, this order on the clusters is not just quasi-linear: it is a *strict linear order*, as required in our definition of a priority base. Accordingly, we define the set $\mathscr{P}^\bullet$ as the set of all clusters. Moreover, we let the order of greater priority run in the upward direction of the cluster order. (This choice of direction is just a convention - but one has to pay attention to it in the following arguments.)

We are now ready to prove our main statement, for all worlds y, x:

$$y \preceq x \quad \text{iff} \quad y \preceq^{OT}_{\mathscr{P}^\bullet} x.$$

(⇒). Assume that $y \preceq x$. We have to show that

$$\forall P \in \mathscr{P}^\bullet (Px \leftrightarrow Py) \vee \exists P' \in \mathscr{P}^\bullet (\forall P < P' (Px \leftrightarrow Py) \wedge (P'x \wedge \neg P'y)).$$

To see this, note that by Lemma 4.2.3, $C_y \trianglelefteq C_x$. Then we distinguish two cases. If $C_y = C_x$, then x, y share this 'property' and no other, and hence the left disjunct holds. If $C_y \neq C_x$, then $C_y \triangleleft C_x$ (by linearity), and then $x \in C_x$ and $y \notin C_x$, and therefore, the right disjunct holds.

($\Leftarrow$). Now assume that $y \preceq^{OT}_{\mathscr{P}^\bullet} x$. There are again two cases. First, let x, y share the same 'properties' in $\mathscr{P}^\bullet$, i.e. $Px \leftrightarrow Py$. Then in particular, $x, y \in C_x$, and hence $y \preceq x$. Next let there be some $P' \in \mathscr{P}^\bullet$ with $P'x \wedge \neg P'y$, while x, y share the same properties $P < P'$. Since $P'x$, we must have $P' = C_x$, while $C_x \neq C_y$. Since we have $\forall P < P'(Px \leftrightarrow Py)$, we conclude that $\neg(C_x \lhd C_y)$. Therefore $C_y \trianglelefteq C_x$, and hence by Lemma 4.2.3, $y \preceq x$. □

Interestingly, the relevant propositions constructed in this argument are all *mutually disjoint*. This may not be the most obvious scenario from the Optimality Theory perspective, but as we shall see later, it is a technically convenient setting, which involves no loss of generality.

Alternative definitions of world orderings

Another appealing definition of preference over alternatives from a priority sequence is called *best-out ordering*. It was used in [BCD+93] as an alternative way of priority-based handling of inconsistent set of classical formulas:

$$y \preceq^{best}_{\mathscr{P}} x ::= \forall P \in \mathscr{P}(Px \wedge Py) \vee \exists P'(\forall P < P'(Px \wedge Py) \wedge (P'x \wedge \neg P'y)).$$

The best-out-format is similar to the OT-definition, except that instead of having the equivalence $(Px \leftrightarrow Py)$ it only requires a conjunction $(Px \wedge Py)$. Intuitively, this means that only the positive cases matter when deriving a preference order from a priority base. Looking at Example 4.2.1 again, we get $d_2 \preceq d_1$ and $d_1 \preceq d_2$, d_1 and d_2 are equally preferable for Alice, because after observing that $\neg Q(d_1)$ and $\neg Q(d_2)$, she won't consider N at all. While these alternatives are interesting, we now move to a slightly more general comparative perspective.

Preference merge and partial order

In what follows, we will consider another formal approach for deriving preference from a priority base, proposed in [ARS02]. The ideas here work differently, but as we shall see, it is a natural generalization of our $\preceq^{OT}_{\mathscr{P}}$ in two ways:

- one merges arbitrary given relations on objects, not just those given by propositions or properties.
- the definition for the merged preference works with partial orders on sets of propositions, which includes our linear priority orders as a special case.

First, we briefly review the basic notions in [ARS02]. A preference relation is any reflexive transitive relation ('pre-order'). Suppose there is a family of such preference relations $(R_x)_{x \in V}$, all on the same set W, V is a set of variables. A question arising in many settings, from social choice theory to 'belief merge' in belief revision theory, is how to combine these relations into a single relation on

the same set W. In particular, the given preferences can originate from different criteria that we wish to combine according to their importance - as in 'many-dimensional decision problems'. But the initial orders do not come as a mere set. We need further structure to arrive at a plausible notion of merge. The core notion here is a *priority graph*, defined in [ARS02] as follows:

4.2.4. Definition. A *priority graph* is a tuple $(N, <, v)$ where N is a set of nodes, $<$ is a strict partial order on N (the 'priority relation') and v is a function from N to a set of variables standing for the given binary relations to be merged. The set N may be infinite, though we will mainly look at finite cases.

A priority graph may be viewed as an ordering of variables for 'input relations'. Some variables may be represented several times in the ordering, simply by repeating their occurrences in the priority graph. Now, any priority graph denotes the following operator on the given preference relations:

4.2.5. Definition. The V-ary operator $\circ$ denoted by the priority graph $(N, <, v)$ is given by

$$m \circ ((R_x)x \in V)n \Longleftrightarrow \forall i \in N.(mR_{v(i)}n \vee \exists j \in N.(j < i \wedge mR^{<}_{v(j)}n))$$

where $V = v[N]$, the set of variables that occur in the graph, and $R^{<}_{v(j)}$ is the strict version of the partial order $R_{v(j)}$.

The operator $\circ$ takes a set of preference relations $(R_x)_{x \in V}$ and returns a single one. The concrete intuition behind Definition 4.2.5 is this:

> $m \preceq_G n$ if for all separate relations $R_{v(i)}$: either $mR_{v(i)}n$ or if n fails this 'test' with respect to m, it 'compensates' for this failure by doing better than m on some more important test, i.e. some relation $R^{<}_{v(j)}$ holds with $j < i$, i.e. with higher priority in the graph.

To see how all this works, we look at the following example from [ARS02]:

4.2.6. Example. The priority graph $g_1 = (N, <, v)$ has $N = \{1, 2, 3\}$ with $1 < 2$ and $1 < 3$ and $v(1) = y$, $v(2) = x$ and $v(3) = y$. See Figure 4.1.

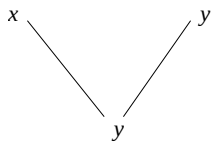

Figure 4.1: Priority graph

Here lower down means higher priority. This graph denotes a binary operator since there are only two distinct variables. It takes two preference relations, say

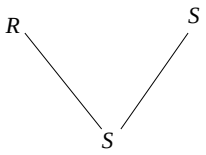

Figure 4.2: Representations of preference relations

R and S, and returns one which represents their combination with the given priority. Thus, if σ_1 is the operator denoted by the graph, then $\sigma_1(R, S)$ is the following prioritized combination of R and S, see Figure 4.2.

Working out what this means by definition 4.2.5, we obtain the relation $\sigma_1(R, S) = (R \cap S) \cup S^{<}$.

To understand further how the ARS system works, note that the above relation $(R \cap S) \cup S^{<}$ is exactly the same as that induced by the simpler linear priority graph shown in Figure 4.3.

Figure 4.3: Linear graph

This is no coincidence. There is a rather interesting algebraic structure behind all this. In particular, [ARS02] proved that every merge operation defined by a priority graph can be derived as an algebraic composition of the following two basic relations:

(i) $(R \cap S) \cup S^{<}$,

(ii) $R \cap S$.

Interestingly, the priority graph inducing the second operation is not linear, but a simple disjoint union, and hence a partial order:

Figure 4.4: Partial order

It is easy to see that by definition 4.2.5, this works out to the intersection of R and S. Indeed, this suggests a much more general observation showing how operations on priority graphs affect the merged outcomes. It can be proved by simple inspection of the above definition:

4.2.7. FACT. For any two priority graphs $\mathscr{P}$ and $\mathscr{P}'$, the following equivalence holds

$$y \preceq_{\mathscr{P} \uplus \mathscr{P}'} x \quad \text{iff} \quad (y \preceq_{\mathscr{P}} x) \,\text{and}\, (y \preceq_{\mathscr{P}'} x),$$

where $\uplus$ denotes the *disjoint union* of the two graphs.[5]

Comparing *ARS* to *OT*

The *ARS* way of thinking matches well with the *OT*-style definition for $\preceq^{OT}$. The latter worked with an ordered set of propositions rather than relations R_i. But it can easily be recast in the latter manner. We merely associate each proposition A with an ordering relation $\preceq(A)$ derived as follows:

$$y \preceq (A)x \quad \text{iff} \quad (Ay \to Ax) \vee (\neg Ay \wedge Ax).$$

This is precisely the sort of world ordering encountered in belief revision when a signal comes in that A is the case (cf. [Rot07], [Ben07a]). Indeed, it is completely interchangeable whether we talk about propositions A or relations $\preceq(A)$: both contain the same information. But the relational format is more general, since not all given object orders need to be generated from propositions in this simple manner. Also, intuitively, the priority order of propositions in the *OT*-format corresponds to the order of relations in the priority graph (where we will disregard the issue of repeated occurrences of variables, which would correspond to repeated occurrences of the same property in a priority sequence).

We can now write the *ARS*-definition as it applies to our proposition orders:

$$y \preceq^{ARS}_{\mathscr{P}} x ::= \forall P \in \mathscr{P}((Py \to Px) \vee \exists P' < P(P'x \wedge \neg P'y)).$$

For a contrast, recall the definition of $\preceq^{OT}_{\mathscr{P}}$:

$$y \preceq^{OT}_{\mathscr{P}} x ::= \forall P \in \mathscr{P}(Px \leftrightarrow Py) \vee \exists P' \in \mathscr{P}(\forall P < P'(Px \leftrightarrow Py) \wedge (P'x \wedge \neg P'y)).$$

Syntactically, this looks quite different, with an inversion in quantifier scope. But actually, the following equivalence result holds:

4.2.8. THEOREM. *For any finite linearly ordered set of propositions $\mathscr{P}$,*

$$y \preceq^{OT}_{\mathscr{P}} x \quad \textit{iff} \quad y \preceq^{ARS}_{\mathscr{P}^*} x \quad \textit{for all worlds} \quad x, y,$$

where $\mathscr{P}^*$ is the priority graph derived from $\mathscr{P}$ by replacing each proposition A by its relation $\preceq(A)$ and keeping the old order from $\mathscr{P}$.

[5] In fact, this is one of the axioms of the graph calculus studied in [Gir08], it is formulated as $\langle G_1 \uplus G_2 \rangle s \leftrightarrow \langle G_1 \rangle s \cap \langle G_2 \rangle s$ there (s is a nominal).

Proof. ($\Rightarrow$). Let $y \preceq^{OT}_{\mathscr{P}} x$. Suppose $y \preceq^{ARS}_{\mathscr{P}^*} x$ does not hold. Then we have

$$\exists P \in \mathscr{P}^*((Py \wedge \neg Px) \wedge \forall P' < P(P'x \rightarrow P'y)).$$

Let P^* be such a P. Notice that $\neg(P^*x \leftrightarrow P^*y)$. By $y \preceq^{OT}_{\mathscr{P}} x$, let $P^\bullet$ be the smallest $P \in \mathscr{P}^*$, s.t. $P^\bullet x \wedge \neg P^\bullet y$. Then $P^\bullet$ cannot come before P^*, since we have $\forall P' < P^*(P'x \rightarrow P'y)$. The only possible case is then $P^\bullet \geq P^*$. Here $P^\bullet = P^*$ leads to a contradiction, as $P^\bullet x$ but $\neg P^* x$. But if $P^\bullet$ comes after P^*, then by the OT-definition, $P^*x \leftrightarrow P^*y$, and again we get a contradiction.

($\Leftarrow$). Let $y \preceq^{ARS}_{\mathscr{P}^*} x$. Suppose it is not the case that $y \preceq^{OT}_{\mathscr{P}} x$, then we have

$$\exists P \in \mathscr{P}\neg(Px \leftrightarrow Py) \wedge \forall P'(\exists P < P'\neg(Px \leftrightarrow Py) \vee (P'x \rightarrow P'y))$$

From $\exists P \in \mathscr{P}\neg(Px \leftrightarrow Py)$, without loss of generality, take $P^* = P$, the smallest $P \in \mathscr{P}$ where $\neg(P^*x \leftrightarrow P^*y)$. Applying the conjunct $\forall P'(\exists P < P'\neg(Px \leftrightarrow Py) \vee (P'x \rightarrow P'y))$ to P^*, which was chosen to be smallest with the non-equivalence, we get that $\neg(P^*y \rightarrow P^*x)$, i.e. $P^*y \wedge \neg P^*x$. Applying the ARS-definition to P^*, we have that $(P^*y \rightarrow P^*x) \vee \exists P' < P^*(P'x \wedge \neg P'y)$. Here the second disjunct does not hold because of P^*'s minimality for equivalence failure. But the first disjunct does not hold either: $(P^*y \rightarrow P^*x)$ cannot occur since $P^*y \wedge \neg P^*x$. So we conclude that $\neg(y \preceq^{ARS} x)$, a contradiction, and we are done. □

The ARS-definition also applies to situations in which the order of $\mathscr{P}$ is partial, and/or infinite, so we can think of it as a natural generalization of our earlier OT-definition.

More on partial orders and pre-orders

There are many technical results in [ARS02], including an algebraic axiomatization of merge operations and a characterization of priority graph-based merge with respect to some conditions from social choice theory. But our reason for considering this system is simply this: the base preference relations over possible worlds considered in Chapter 2 are reflexive, transitive pre-orders, rather than quasi-linear ones. Moreover, in the context of dynamic relation transformations, reflexivity and transitivity were preserved (for a proof, see Chapter 2), but not in general connectedness. We will return to this issue later, but for the moment, we state a representation result comparable to Theorem 4.2.2, which works for object pre-orders and partial graph order, thereby generalizing our earlier case of quasi-linear object order and strictly linear constraint order.

4.2.9. Theorem. *Let $\mathcal{M} = (W, \preceq, V)$ be an ordinary object model. Then the following two statements are equivalent:*

(a) there is a structured model $\mathcal{M}' = (W, \preceq, V, (\mathscr{P}, <))$*, with* $(\mathscr{P}, <)$ *a priority graph over given propositional relations* $\preceq(A)$ *such that*

$$y \preceq x \quad \textit{iff} \quad y \preceq_{\mathscr{P}}^{ARS} x \quad \textit{for all} \quad x, y \in W.$$

(b) $y \preceq x$ *is a reflexive transitive pre-order.*

Proof. (a)⇒ (b). It is easy to see from the earlier definition that all relations $\preceq_{\mathscr{P}}^{ARS}$ generated by priority graphs decorated with pre-orders must be reflexive and transitive.
Conversely, consider the direction (b)⇒ (a). Similarly to the proof of Theorem 4.2.2, we can define the set of all clusters C_x. We define an order over clusters in the same way:

$$C \trianglelefteq C' \quad \text{if} \quad \exists y \in C, \exists x \in C' : y \preceq x.$$

Lemma 4.2.3 holds for pre-orders as well, since its proof did not rely on the quasi-linearity of the object order. Indeed, there is again an 'improvement': the cluster ordering becomes a *partial order*: i.e. two clusters which mutually precede each other must be the same. (The latter property is not true for pre-orders in general, and not even for quasi-linear orders.)

As we have seen in the preceding, each 'cluster proposition' P is associated with an ordering relation $\preceq(P)$, so the partial order of clusters gives us a partially ordered priority graph. Again we need to show that the relation induced by this matches up with the given one, that is:

$$y \preceq x \quad \text{iff} \quad y \preceq_{\mathscr{P}}^{ARS} x.$$

(⇒). Assume that $y \preceq x$. We need to show that $y \preceq_{\mathscr{P}}^{ARS} x$, i.e., $\forall P \in \mathscr{P}((Py \rightarrow Px) \vee \exists P' < P \wedge P'x \wedge \neg P'y))$. So, consider any cluster proposition P, or its associated relation. If $Py \rightarrow Px$, we are done. So suppose $Py \wedge \neg Px$. Then we must have $P = C_y$, since as before, our cluster propositions form a disjoint partition. Moreover, we have $C_y \trianglelefteq C_x$ by our Lemma applied to $y \preceq x$, and since $C_y \neq C_x$ (they are disjoint since x is not in C_y), we get $C_y \triangleleft C_x$. It is easy to see that C_x is the 'compensating' P' for x that we need to verify the second disjunct in the *ARS*-definition.
(⇐). Assume that $y \preceq_{\mathscr{P}}^{ARS} x$. Consider the predicate $P = C_y$, for which clearly Py holds. There are two cases. First assume that Px. Since $P(= C_y)$ is a cluster, we have that $y \preceq x$. Next, assume that not Px. By the 'compensation clause' of $y \preceq^{ARS} x$, $\exists P' < P : P'x \wedge \neg P'y$. Clearly, this P' can only be C_x, and hence we have $C_y \triangleleft C_x$, $C_y \trianglelefteq C_x$, and by Lemma 4.2.3, we get $y \preceq x$. □

We will continue our discussion of the *ARS*-format in later sections. For now we just make one simple but useful observation:

4.2.10. FACT. Let $\mathscr{P}; A$ be a set of ordered priorities with a priority A at the end, then we have

$$y \preceq^{ARS}_{\mathscr{P};A} x \Rightarrow y \preceq^{ARS}_{\mathscr{P}} x.$$

Proof. If there is a predicate P' in $\mathscr{P}; A$ such that $P'y \wedge \neg P'x$, then by the *ARS*-definition, there must be a 'higher compensating predicate in $\mathscr{P}; A$, but given that A comes last, this compensation can only happen inside $\mathscr{P}$. □

4.3 Going from world preference to propositional priorities

Now let us take the opposite perspective to that of the preceding Section 4.2. This time, a primitive order over worlds is given, and we would like to lift it to an order over propositions, so that we can compare sets of possible worlds.

This scenario, too, occurs in many places in the literature, with various interpretations of the basic relation $y \preceq x$. It is interpreted as 'x is as least as normal (or typical) as y' in [Bou94] on conditional and default reasoning, as 'x at least as preferred or desirable as y' in [DSW91], as 'x is no more remote from actuality than y' in [Lew73] on counterfactuals, and as 'x is as likely as y' in [Hal97] on qualitative reasoning with probability. In all these settings, it makes sense to extend the given order on worlds to an order of propositions P, Q. For instance, in real life, students may have preferences concerning courses, but they need to also form an order over kinds of courses, say theoretical versus practical, i.e. over sets of individual courses. Likewise, we may have preferences regarding individual commodities, but we often need a preference over sets of them. And similar aggregation scenarios are abundant in social choice theory, for which an extensive survey is [BRP01].

Quantifier lifts

One obvious way of lifting world orders $x \preceq y$ to proposition or set orders $P \preceq Q$ uses definitional schemas that can be classified by the quantifiers which they involve. As has been observed by many authors (cf. [BRG07]), there are four obvious two-quantifier combinations:

$$\forall x \in P \forall y \in Q : x \preceq y; \quad \forall x \in P \exists y \in Q : x \preceq y;$$
$$\exists x \in P \forall y \in Q : x \preceq y; \quad \exists x \in P \exists y \in Q : x \preceq y.$$

One can argue for any of these. [BOR06] claims that $\forall\forall$ is the notion of 'preference' intended by von Wright in his seminal work on preference logic ([Wri63]) and provides an axiomatization.[6] But the tradition is much older, and (modal)

[6] [BOR06] need to assume quasi-linearity of the world preference relation to define the lifted relation within their modal language.

logics for preference relations over sets of possible worlds have been considered by [Lew73], [Bou94] and [Hal97], and other authors. In particular, [Hal97] studied the above combination of $\forall\exists$, defined more precisely as follows:

4.3.1. DEFINITION. Let $(W, \preceq, V, (\mathscr{P}, <))$ be any structured model. For $X, Y \in \mathscr{P}$, we define $Y \trianglelefteq^{\forall\exists} X$ if for all $y \in Y$, there exists some $x \in X$ with $y \preceq x$.

As usual, we define the strict variant $Y \triangleleft X$ as '$Y \trianglelefteq^{\forall\exists} X$ and not $X \trianglelefteq^{\forall\exists} Y$'. Similarly, we can define all the other quantifier combinations.

As we have seen in Chapter 2, some of these combinations can be expressed directly in a modal language for the models $\mathcal{M} = (W, \preceq, V)$ by combining a betterness modality with a universal modality.[7] As for the $\forall\exists$-preference, it can be defined in such a modal language as follows:

$$Pref^{\forall\exists}(\varphi, \psi) ::= U(\psi \rightarrow \langle bett \rangle \varphi).$$

This says that, for any ψ-world in the model, there exists a better φ-world. Once again, this 'majorization' is one very natural way of comparing sets of possible worlds - and it has counterparts in many other areas which use derived orders on powerset domains. In particular, [Hal97] took this definition (with an interpretation of 'relative likelihood' between propositions) and gave a complete logic for the case in which the basic order on W is a pre-order. It is also well-known that Lewis gave a complete logic for preference relations over propositions in his study of counterfactuals in [Lew73], where the given order on W is quasi-linear. In what follows, we side-step these completeness results,[8] but raise a few more semantic issues, closer to understanding the lifting phenomenon per se.

More on $\forall\exists$-preference

A natural question to ask is: Which lift is 'the right one'? This is hard to say, and the literature has never converged on any unique proposal. There are some obvious necessary conditions, of course, such as the following form of 'conservatism':

`Extension rule`: For all $x, y \in X$, $\{y\} \trianglelefteq \{x\}$ iff $y \preceq x$.

But this does not constrain our lifts very much, since all four quantifier combinations satisfy it. We will not explore further constraints here. Instead, we concentrate on one particular lift, and try to understand better how it works. One question that comes to mind immediately is this: Can the properties of

[7] [BOR06] shows that the standard modal language plus universal modality is not sufficiently expressive to define the intended meaning of $\forall\forall$ or $\exists\forall$. The language has to be extended further by a strict preference operator $[bett^s]$.

[8] See however our discussion in Section 3.7.

an underlying preference on worlds be preserved when it is lifted to the level of propositions? In particular, consider reflexivity and transitivity that we assumed for preference in Chapter 2. Can we show $\preceq^{\forall\exists}(\varphi, \psi)$ has these two properties? We can even prove something stronger:

4.3.2. Fact. Reflexivity and transitivity of the relation $\preceq$ are preserved in the lifted relation $\preceq^{\forall\exists}$, but also vice versa.

Proof. *Reflexivity.* To show that $\preceq^{\forall\exists}(X, X)$, by Definition 4.3.1, we need that $\forall x \in X \exists y \in X : x \preceq y$. Since we have $x \preceq x$, take y to be x, and we get the result.

In the other direction, we take $X = \{x\}$. Then apply $\preceq^{\forall\exists}(X, X)$ to it to get $\forall x \in X \exists x \in X : x \preceq x$. Since x is the only element of X, we get $x \preceq x$.

Transitivity. Assume that $\preceq^{\forall\exists}(X, Y)$ and $\preceq^{\forall\exists}(Y, Z)$. We show that $\preceq^{\forall\exists}(X, Z)$. By Definition 4.3.1, this means we have $\forall x \in X \exists y \in U : x \preceq y$ and $\forall y \in Y \exists z \in Z : y \preceq z$. Then by transitivity of the base relation, we have that $\forall x \in X \exists z (x \preceq z$, and this is precisely $\preceq^{\forall\exists}(X, Z)$.

In the other direction, let $x \preceq y$ and $y \preceq z$. Take $X = \{x\}, Y = \{y\}$ and $Z = \{z\}$. Applying $\preceq^{\forall\exists}$, we see that $X \preceq Y$ and $Y \preceq Z$, and hence by transitivity for sets, $X \preceq Z$. Unpacking this, we see that we must have $x \preceq z$. □

Likewise, we can prove that if $\preceq^{\forall\exists}$ is quasi-linear, then so is $\preceq$. But the converse direction does not hold.

Besides the three properties mentioned, many others make sense. In fact, the preceding simple argument suggests a more extensive correspondence between relational properties for individual orderings and their set liftings, which we do not pursue here.

Next, staying at the level of propositions, consider an analogue to the representation theorems of the preceding section. Suppose we have a preference relation that is a $\forall\exists$-lift from a base relation over possible worlds. What are necessary and sufficient conditions for being such a relation? The following theorem provides a complete characterization.

4.3.3. Theorem. *A binary relation $\preceq$ over propositions satisfies the following four properties iff it is a $\forall\exists$-lifting of some preference relation over the underlying possible worlds.*

(1) $Y \preceq X \Rightarrow Y \bigcap Z \preceq X$ (*left downward monotonicity*)

(2) $Y \preceq X \Rightarrow Y \preceq X \bigcup Z$ (*right upward monotonicity*)

(3) $\forall i \in I, Y_i \preceq X \Rightarrow \bigcup_i Y_i \preceq X.$ (*left union property*)

(4) $\{y\} \trianglelefteq \bigcup_i X_i \Rightarrow \{y\} \trianglelefteq X_i$ *for some* $i \in I$. (*right distributivity*)

Proof. ($\Leftarrow$). Assume that $\trianglelefteq$ is a $\forall\exists$-lifting. We show that $\trianglelefteq$ has the four properties.

(1). Assume $Y \trianglelefteq X$, i.e., $\forall y \in Y \exists x \in X : y \preceq x$. Since $Y \bigcap Z \subseteq Y$, we also have $\forall y \in Y \bigcap Z \exists x \in X : y \preceq x$, and hence $Y \bigcap Z \trianglelefteq X$.

(2). Assume $\forall y \in Y \exists x \in X : y \preceq x$. Since $X \subseteq X \bigcup Z$, we have $\forall y \in Y \exists x \in X \bigcup Z : y \preceq x$: that is, $Y \trianglelefteq X \bigcup Z$.

(3). Assume that for all $i \in I$, $\forall y \in Y_i \exists x \in X : y \preceq x$. Let $y \in \bigcup_i Y_i$, then for some j: $y \in Y_j$. By the assumption, we have $\forall y \in Y_j \exists x \in X : y \preceq x$, so $\exists x \in X : y \preceq x$. This shows that $\bigcup_i Y_i \trianglelefteq X$.

(4). Assume that $\forall y \in \{y\} \exists x \in \bigcup_i X_i : y \preceq x$. Then there exists some X_i with $\exists x \in X_i : y \preceq x$, that is: $\{y\} \trianglelefteq X_i$ for some $i \in I$.

($\Rightarrow$). Going in the opposite direction, we first define an object ordering

$$y \preceq x \quad \text{iff} \quad \{y\} \trianglelefteq \{x\}. \quad (\dagger)$$

Next, given any primitive relation $Y \trianglelefteq X$ with the above four properties, we show that we always have

$$Y \trianglelefteq X \quad \text{iff} \quad Y \trianglelefteq^{\forall\exists} X.$$

where the latter relation is the lift of the just-defined object ordering.
($\Rightarrow$). Assume that $Y \trianglelefteq X$. For any $y \in Y$, $\{y\} \subseteq Y$ by reflexivity. Then, by Property (1) we get $\{y\} \trianglelefteq X$. But then also $\{y\} \trianglelefteq \bigcup_{x \in X} \{x\}$, as $X = \bigcup_{x \in X} \{x\}$. By Property (4), there exists some $x \in X$ with $\{y\} \trianglelefteq \{x\}$, and hence, by Definition ($\dagger$), $y \preceq x$. This shows that, for any $y \in Y$, there exists some $x \in X$ s.t. $y \preceq x$, which is to say that $Y \trianglelefteq^{\forall\exists} X$.

($\Leftarrow$). Assume that $\forall y \in Y \exists x \in X : y \preceq x$. By definition ($\dagger$), $y \preceq x$ is equivalent to $\{y\} \trianglelefteq \{x\}$. Since $\{x\} \subseteq X$, by Property (2) we get that $\{y\} \trianglelefteq X$. Thus, for any $y \in Y$, $\{y\} \trianglelefteq X$. By Property (3) then, $\bigcup_{y \in Y} \{y\} \trianglelefteq X$, and this is just $Y \trianglelefteq X$.[9] $\square$

Finally, we ask how orderings of propositions produced by set lifting relate to the priority orderings which were central in Chapter 3 and Section 4.2. Intuitively, there need not be any strong connection here, since priority ordering is about relative importance, rather than preference. Nevertheless, in some special cases, we can say more. We have some results of this sort on the the $\trianglelefteq^{\forall\exists}$-lifting, but instead, we cite an observation from Chapter 3 which is relevant here. Priority order and lifted object order can coincide when we work with special sets of worlds. Here is how:

[9] [Hal97] gave a complete logic for $Pref^{\forall\exists}$, whose axiomatization looks different from our characterization. But one should be able to show they are essentially equivalent.

4.3.4. Definition. A set X is *upward closed* if

$$\forall x, y \in X(y \in X \wedge y \preceq x \rightarrow x \in X).$$

4.3.5. Fact. Consider only sets X that are upward closed. We define

$$y \preceq x \quad \text{iff} \quad \forall X(x \in X \leftrightarrow y \in X) \vee \exists X(x \in X \wedge y \notin X).$$

Then the $\trianglelefteq^{\forall\exists}$-lifting of this object ordering becomes equivalent to set inclusion, and the latter is equivalent to the priority sequence:

$$X \subseteq Y \Leftrightarrow X \gg Y.$$

Much more general questions arise here about connections when transformations are repeated:

- Given a priority order $(\mathscr{P}, <)$ and its induced world order $\preceq^{ARS}_{\mathscr{P}}$(or $\preceq^{OT}_{\mathscr{P}}$), when can $<$ on $\mathscr{P}$ be retrieved as a quantifier lift of $\preceq^{ARS}_{\mathscr{P}}$?
- Given a world order $(W, \preceq)$, and some lift, say $\trianglelefteq^{\forall\exists}$, used as a priority order on the powerset $\mathcal{P}(W)$, when can the relation $\preceq$ on W be retrieved as the derived order of $\trianglelefteq^{\forall\exists}$?

Answering these questions would show us further connections between the two levels of worlds and propositions, when both relative importance and preference are involved in lifting and deriving. We will not pursue these matters here - but there is certainly more harmony than what we have uncovered so far.

4.4 Dynamics at two levels

Dynamics has been one of the core issues investigated in the previous two chapters. We have modeled changes in the betterness relation over possible worlds in Chapter 2. And we also considered possible changes in priority sequences in Chapter 3. As we pointed out after introducing the structured models in Section 4.1, we want to relate the changes at the two levels in a systematic manner.

Relation transformers at the world level

In Chapter 2 betterness relations over possible worlds are the locus of dynamics. New information or other triggers come in which rearrange this order. We recapitulate a few concrete operations from Chapter 2.

The simplest operation was $Cut(A)$. It cuts only the accessibility links between worlds in A and $\neg A$, but keeps all possible worlds around.[10] The following

[10]This is different from eliminative update for public announcement, where worlds may disappear. With preference change, belief revision, or even information update for memory-bounded agents, it is reasonable to keep all possible worlds, as shown in [BL07], [Ben07a], and [Liu07].

definition of $Cut(A)$ is stated using some self-explanatory notation from propositional dynamic logic:

4.4.1. DEFINITION. For any relation R and proposition A, the new relation $Cut(A)(R)$ is defined as:

$$Cut(A)(R) ::= (?A; R; ?A) \cup (?\neg A; R; ?\neg A).^{11}$$

Next, 'suggesting A' (written as $\sharp A$) was the main action considered in Chapter 2, changing (preference) relations in the following manner:

4.4.2. DEFINITION. For any relation R and proposition A, the new relation $\sharp A(R)$ is defined as:

$$\sharp A(R) ::= (?A; R; ?A) \cup (?\neg A; R; ?\neg A) \cup (?\neg A; R; ?A).$$

Thus, in the updated models no $\neg A$-worlds are preferable to A-worlds. We can also define the suggestion relation by

$$Cut(A)(R) \cup (?\neg A; R; ?A).$$

In fact, $Cut(A)$ is a basic operation, and it will return below. One slightly more complex relevant operation is 'upgrade with A', written as $\Uparrow A$:

4.4.3. DEFINITION. The new relation $\Uparrow A(R)$ is defined as:

$$\Uparrow A(R) ::= Cut(A)(R) \cup (?\neg A; \top; ?A).$$

After the new information A has been incorporated, the upgrade places all A-worlds on top of all $\neg A$-worlds, keeping all other comparisons the same. This time, besides $Cut(A)$ as before, new links may be added by the disjunct $(?\neg A; \top; ?A)$ with the *universal relation* $\top$. Alternatively, going back to Section 4.2, $\Uparrow A(R)$ can also be defined as

$$Cut(A)(R) \cup \preceq (A)(R).$$

Here, it is important to note that the above definitions make the ordering over possible worlds hold between whole 'zones' of the model given by propositions. Worlds which satisfy exactly the same propositions behave the same. Thus, we are ordering 'kinds of worlds' (through a partition of the domain of worlds), rather than worlds per se. Keeping this way of thinking in mind helps understand many technicalities in what follows.

[11] Note that the link-cutting operation is found as *agenda expansion* in [BRG07] and *PDL test action* in [HLP00].

We have already used a simple program fragment of the PDL language to define new relations over possible worlds after some definable operation has taken place. Let us look at this more generally. The basic elements that build up the new relations are:

$$?\varphi \mid ; \mid \cup \mid R \mid \top.$$

These are the standard PDL operations. $?\varphi$ is a test, while ; and $\cup$ denote sequential composition and choice, respectively. R is the given input relation, treated as an atom, and the constant $\top$ denotes the universal relation that holds everywhere. The following fact provides a useful 'normal form':

4.4.4. Fact. Every PDL operation in the above style has a definition as a union of finite 'trace expressions' of the form '$?A_1; \{R, \top\}; ?A_2; \{R, \top\}; ...$', where $\{R, \top\}$ means either R or $\top$.

Proof. To turn arbitrary program expressions into normal form, we apply the following equivalences that drive unions outwards:

$$\begin{array}{lcl} (S \cup T); U & = & S; U \cup T; U, \\ S; (T \cup U) & = & S; T \cup S; U, \end{array}$$

where S, T, U, and V are of any form of $?\varphi$, or R or $\top$ from the PDL language. This may still leave us with sequences of tests, instead of single ones. But the former can be contracted using the valid identity

$$?A_1; ?A_2 \quad = \quad ?(A_1 \wedge A_2). \qquad \square$$

In fact, our definitions of the operations $Cut(A)$, $\Uparrow A$, and $\sharp A$ were already in normal form. But in principle, many relations can be defined in a PDL format, covering a large space of possible relation transformers. While some of these make intuitive sense, others are just mathematical curiosities. We will return to this issue later on.

Propositional level transformers of priority orders

At the level of linearly ordered finite sets of priorities $(\mathscr{P}, <)$, some natural operations have been considered in Chapter 3. These were:

- $[^+A]$ adds A to the right of the sequence,
- $[A^+]$ adds A to the left,
- $[-]$ drops the last element of the sequence,
- $[i \leftrightarrow i+1]$ interchanges the i-th and i+1-th elements.

We have shown that any one of the first two operations, plus the last two are sufficient to make any changes to a finite ordered priority sequence. Some operations are considered in [ARS02] as well, now of course on priority graphs, which generalize finite sequences. Examples are taking the disjoint union of two relations, deleting a node from a priority graph, and putting a new link j below i if this does not change the down-set of i. To keep things simple, in what follows, we consider only graph analogues of the preceding 'postfixing' and 'prefixing' operators $[^+A]$ and $[A^+]$, which we will write as $\mathscr{P}; A$ and $A; \mathscr{P}$. In the special case of flat sets $\mathscr{P}$, both collapse to one operation $\mathscr{P} + A$.

Relating dynamics at the two levels

Having reviewed some basic operations at the two levels, let us now try to find systematic correspondences between them. First, we state two trivial but general observations that are easily obtained from the definitions and representation theorems in the previous sections. These work globally in that we do not need to identify which specific dynamic changes have taken place. For convenience, we state several results in the *OT*-setting, but our results hold for *ARS*-style pre-orders as well in general. We start by taking object-level transformers to propositional ones.

4.4.5. Claim. *Given the OT-definition, if a relation change over possible worlds models respects quasi-linear order, then there exists a corresponding change on the set of ordered propositions.*

Proof. Assume the old relation is R_1 and after a change it becomes R_2. Since R_1 and R_2 are both quasi-linear orders, Theorem 4.2.2, gives corresponding priority sequences $\mathscr{P}_1$ and $\mathscr{P}_2$. This is the propositional change we are after. □

Of more interest is the question whether the induced change between $\mathscr{P}_1$ and $\mathscr{P}_2$ can be *defined* using the given one from R_1 to R_2. We discuss this issue later on. For now, here is the converse general observation, which is even more trivial.

4.4.6. Claim. *Given the OT-definition, if a relation change over propositions respects the linear order, then there is a corresponding change in preference over possible worlds.*

This follows at once from the earlier definitions.

Uniform definable connections

Next, we consider more uniform connections between transformations at the two levels. We start with the following notion about relation-transforming functions.

4.4.7. Definition. Let F: $(\mathscr{P}, A) \to \mathscr{P}'$, where $\mathscr{P}$ and $\mathscr{P}'$ are set of propositions, and A is a new proposition. Let σ: $(\preceq, A) \to \preceq'$, where $\preceq$ and $\preceq'$ are relations over possible worlds, and A is a new proposition. We say that *the map F induces the map σ*, given a definition of deriving object preferences from propositions, if, for any set of propositions $\mathscr{P}$ and new proposition A, we have

$$\sigma(\preceq_{\mathscr{P}}, A) = \preceq_{F(\mathscr{P},A)} .$$

We start our discussion of such connections with a simplest case, viz. adding a new proposition to a flat priority set. Recall the $(*)$-definition for inducing object order in Section 4.2.

4.4.8. Fact. Given the $(*)$-definition, taking a suggestion A given some relation over possible worlds is induced by the following operation at the propositional level: adding a new proposition A to a flat $\mathscr{P}$. More precisely, the following diagram commutes:

$$\begin{array}{ccc} \langle W, \mathscr{P}\rangle & \xrightarrow{+A} & \langle W, \mathscr{P} \cup A\rangle \\ {\scriptstyle *}\downarrow & & \downarrow{\scriptstyle *} \\ \langle W, \preceq\rangle & \xrightarrow{\sharp A} & \langle W, \sharp A(\preceq)\rangle \end{array}$$

Proof. We need to prove the following equivalence:

$$y \preceq^*_{\mathscr{P}+A} x \quad \text{iff} \quad y \sharp A(\preceq^*_{\mathscr{P}}) x.$$

$(\Leftarrow)$. We know that after $\sharp A$, the relation between y and x can be expressed as:

$$\sharp A(\preceq^*_{\mathscr{P}}) ::= (?A; \preceq^*_{\mathscr{P}}; ?A) \cup (?\neg A; \preceq^*_{\mathscr{P}}; ?\neg A) \cup (?\neg A; \preceq^*_{\mathscr{P}}; ?A)$$

In terms of a relation between arbitrary worlds x and y, the above three cases give the implication $Ay \to Ax$. By $y \preceq^*_{\mathscr{P}} x$, we also have that $\forall P \in \mathscr{P}$: $Py \to Px$. Hence $\forall P \in \mathscr{P} + A$: $Py \to Px$: i.e., $y \preceq^*_{\mathscr{P}+A} x$.
$(\Rightarrow)$. Assume that $y \preceq^*_{\mathscr{P}+A} x$, i.e., $\forall P \in \mathscr{P} + A$: $y \in P \to x \in P$. In particular, it cannot the case that $y \in A \wedge x \notin A$. Thus, out of all pairs in the given relation R, those satisfying $(?A; \preceq^*_{\mathscr{P}}; ?\neg A)$ can no longer occur. This is precisely how we defined the relation $y \sharp A(\preceq^*_{\mathscr{P}}) x$. □

Simple as it is, this argument shows that natural operations at both levels can be tightly correlated.

Next we consider the case of an *ordered* set $(\mathscr{P}, <)$, where a new proposition A is added in front. The dynamics at the two levels is correlated as follows:

4.4.9. FACT. Given the OT-definition, upgrade $\Uparrow A$ over possible worlds is induced by the following operation on propositional priority orders: prefixing a new A to an ordered propositional set $(\mathscr{P}, <)$. More precisely, the following diagram commutes:

$$\begin{array}{ccc} \langle W, (\mathscr{P}, <)\rangle & \xrightarrow{A;\mathscr{P}} & \langle W, (A; \mathscr{P}, <)\rangle \\ {\scriptstyle OT}\downarrow & & \downarrow{\scriptstyle OT} \\ \langle W, \preceq\rangle & \xrightarrow{\Uparrow A} & \langle W, \Uparrow A(\preceq)\rangle \end{array}$$

Proof. Again, we have to prove a simple equivalence:

$$y \preceq^{OT}_{A;\mathscr{P}} x \quad \text{iff} \quad y \Uparrow A(\preceq^{OT}_{\mathscr{P}})x.$$

($\Leftarrow$). We know that after the operation $\Uparrow A$, the relation between y and x can be expressed as:

$$\Uparrow A(\preceq^{OT}_{\mathscr{P}}) ::= (?A; \preceq^{OT}_{\mathscr{P}}; ?A) \cup (?\neg A; \preceq^{OT}_{\mathscr{P}}; ?\neg A) \cup (?\neg A; \top; ?A).$$

Call these Cases (a), (b) and (c), respectively. We show that $y \preceq^{OT}_{A;\mathscr{P}} x$, i.e.,

$$\forall P \in A; \mathscr{P}(Px \leftrightarrow Py) \vee \exists P' \in A; \mathscr{P}(\forall P < P'(Px \leftrightarrow Py) \wedge (P'x \wedge \neg P'y)).$$

In Case (a) and (b), the new predicate A in top position does not distinguish the worlds x, y , and hence their order is determined by just that in $\mathscr{P}$. In Case (c), since $(Ax \wedge \neg Ay)$, for any pair of y and x, A is the compensating predicate P' in $A; \mathscr{P}$ that we need for the OT-definition. Thus in all cases, we have $y \preceq^{OT}_{A;\mathscr{P}} x$.

($\Rightarrow$). Assume that $y \preceq^{OT}_{A;\mathscr{P}} x$. Consider the following two cases. (i) For all $P \in A; \mathscr{P}$ $Px \leftrightarrow Py$. In particular then, $Ax \leftrightarrow Ay$, and we get Cases (a) and (b). (ii) There exists some $P' \in A; \mathscr{P}$ such that for all $P < P'(Px \leftrightarrow Py)$ while $(P'x \wedge \neg P'y)$. Then $P' = A$ or $P' \in \mathscr{P}$. If $P' = A$, $Ax \wedge \neg Ay$, and we get Case (c). If $P' \in \mathscr{P}$, then, by the prefixing, $A < P'$, by assumption we have $Ax \leftrightarrow Ay$, and again we get Case (a) and (b). □

We have now proved two results in a similar format, linking operations at the possible world level to operations at the priority level. Actually, one can think of such connections in two ways:

(i) Given any priority-level transformer, we define a matching world-level relation transformer.

(ii) Given any world-level relation transformer, we define a matching priority-level transformer.

As an instance of direction (i), let us consider the natural operation $\mathscr{P};A$ of postfixing a proposition to an ordered propositional set. It turns out that we do not have a very simple corresponding operation at the possible world level. We need some relational algebra beyond the earlier PDL-format, witness the following observation.

4.4.10. Fact. $y \preceq^{ARS}_{\mathscr{P};A} x$ iff $y \prec^{ARS}_{\mathscr{P}} x \vee (y \preceq^{ARS}_{\mathscr{P}} x \wedge (Ay \to Ax))$.

Proof.($\Rightarrow$). Assume that $y \preceq^{ARS}_{\mathscr{P};A} x$. By Fact 4.2.10, this implies that $y \preceq^{ARS}_{\mathscr{P}} x$. Now consider the following two cases:
(i) $Ay \to Ax$. Then the right disjunct $(y \preceq^{ARS}_{\mathscr{P}} x \wedge (Ay \to Ax))$ holds.
(ii) $Ay \wedge \neg Ax$. Recall that $y \prec^{ARS}_{\mathscr{P}} x$ was defined as

$$y \preceq^{ARS}_{\mathscr{P}} x \wedge \neg(x \preceq^{ARS}_{\mathscr{P}} y).$$

We have to prove this conjunction, of which we have the left conjunct already. Suppose that $x \preceq^{ARS}_{\mathscr{P}} y$, we will derive a contradiction. Since $Ay \wedge \neg Ax$, according to the ARS-definition applied to $y \preceq^{ARS}_{\mathscr{P};A} x$, we have $\exists P' \in \mathscr{P};A(P' < A \wedge P'x \wedge \neg P'y)$. Note that $P' \in \mathscr{P}$, and hence we get $\exists P'' \in \mathscr{P}(P'' < P' \wedge P''x \wedge \neg P''y)$. Repeating these two steps, we get an infinite downward sequence (or a finite cycle) of 'compensations' in $\mathscr{P}$, and this contradicts the well-foundedness property of the priority graph.
($\Leftarrow$). Again consider two cases:
(i) $y \prec^{ARS}_{\mathscr{P}} x$. Then $\exists P \in \mathscr{P}$: $Px \wedge \neg Py$. By the definition of $\mathscr{P};A$, any such P satisfies $P < A$. So this P is the compensation in $\mathscr{P};A$. Hence $y \prec^{ARS}_{\mathscr{P};A} x$.
(ii) $y \preceq^{ARS}_{\mathscr{P}} x \wedge (Ay \to Ax)$. We have to show $y \preceq^{ARS}_{\mathscr{P};A} x$ for the extended priority graph $\mathscr{P};A$. Now if $Py, \neg Px$ holds for any $P \in \mathscr{P};A$, then either $P \in \mathscr{P}$ or $P = A$. If $P \in \mathscr{P}$, then it has a compensation P' in the set $\mathscr{P}$ such that $P'x$ and $\neg P'y$. P' is also in $\mathscr{P};A$. Hence $y \preceq^{ARS}_{\mathscr{P};A} x$. If $P = A$, then we would have $Ay, \neg Ax$: but this contradicts $(Ay \to Ax)$. □

Next we illustrate direction (ii) from given object-relation-transformers to priority operations. Consider our 'suggestion' operation $\sharp A$ at the level of possible worlds. First, observe what this does for linearly ordered priorities:

4.4.11. Example. Let $\mathscr{P} = \{P\}$, which is trivially linearly ordered. This makes any P-world more preferable than any $\neg P$-world. After the suggestion A comes in, $\neg A$-worlds can no longer be preferable to A-worlds. This results in what is shown in Figure 4.5.

In particular, looking at worlds s and t, $\neg P \wedge A$ is true in s, and $P \wedge \neg A$ is true in t. But s and t are not comparable in this new model, while in the old model, t was preferable to s (since t was a P-world and s was not). Thus, we lose the connectedness.

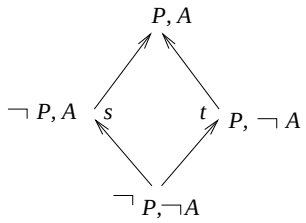

Figure 4.5: Loss of connectedness

Thus, given some relation change at the level of possible worlds, it is by no means the case that there must be a simple corresponding priority operation. As we have just seen:

4.4.12. Fact. Given a linear propositional set $(\mathscr{P}, <)$. $\sharp A$ is not induced by any F that preserves linearity.

Proof. Linear propositional sets induce quasi-linear orders, and we have just seen how suggestions can lose the quasi-linearity. □

However, if we take a partially ordered constraint set, i.e. if we move to priority graphs once more, then we get a positive result.

4.4.13. Fact. The operation $\sharp A$ is induced by the following operation F on priority graphs $\mathscr{P}$:

$$F(A, \mathscr{P}) = (A; \mathscr{P}) \uplus (\mathscr{P}; A).$$

Proof. Note that we take a disjoint union here where the same A occurs twice, but at different positions. We show the following equivalence:

$$y \preceq^{ARS}_{A;\mathscr{P}\uplus\mathscr{P};A} x \quad \text{iff} \quad y\sharp A(\preceq^{ARS}_{\mathscr{P}})x.$$

By Fact 4.2.7, we need to show:

$$y \preceq^{ARS}_{A;\mathscr{P}} x \text{ and } y \preceq^{ARS}_{\mathscr{P};A} x \quad \text{iff} \quad y\sharp A(\preceq^{ARS}_{\mathscr{P}})x.$$

On the left-hand side, by Facts 4.4.9 and 4.4.10, we get

$$y \Uparrow A(\preceq_{\mathscr{P}})x \text{ and } (y \prec_{\mathscr{P}} x \vee (y \preceq_{\mathscr{P}} x \wedge (Ay \to Ax)).$$

Figure 4.6 depicts what happens with the first conjunct of this formula. And Figure 4.7 depicts what happens with the second conjunct.

Clearly, the intersection of these two relations gets us precisely what we had in Example 4.4.11. The partial order on the priority graph allows for intersection, and hence incomparable situations. □

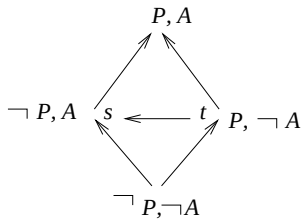

Figure 4.6: A-worlds preferable to $\neg A$-worlds

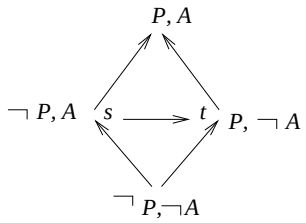

Figure 4.7: P-worlds preferable to $\neg P$-worlds

General formats of definition

These examples raise some general questions. In particular, can every PDL-definable world-level relation transformer which takes some old relation R and new proposition A as input be generated by some simple operation on priority graphs? This is not easy to answer in general, and we will merely provide some discussion. To match our examples so far, we first restrict the PDL-format in the following two aspects:

(i) We only consider atomic tests $?A$, $?\neg A$.

(ii) We only consider normal forms with single occurrences of the relation R. Forms such as $?A; R; R; ?\neg A$ will be disregarded.[12]

Given the above restrictions, we can enumerate all possible Cases:

4.4.14. Fact. There are at most 2^8 basic PDL-transformers over possible worlds.

Proof. Disjuncts in the normal form look like this:

$$\{?A, ?\neg A\}; \{R, \top\}; \{?A, ?\neg A\}$$

where at each position, there are 2 possible options, giving 8 basic cases. To define the new relation, any of these may or may not occur, giving us the exponent. □

For instance, $\Uparrow A(R)$ was a union of the three basic cases

[12] Technically, this restriction makes our operation completely distributive in its R-argument.

$$(?A; R; ?A) \cup (?\neg A; R; ?\neg A) \cup (?\neg A; \top; ?A).$$

As observed before, some of these relation changes can be induced by a change in priorities, some cannot. Moreover, not all preserve the base properties of reflexivity and transitivity. For a counter-example, take $?A; R$, that is: 'if A is true, keep the old relation'. This does not preserve reflexivity, as $\neg A$-worlds have no relations any more. So this relation-transformer cannot be defined even using a partial priority graph. $\sharp A$ does yield reflexive and transitive orders (see our brief proof in Chapter 2), but not always connected ones (see Example 4.4.11), and we have just seen it is not definable by a map on linear priority graphs - though we defined it with one on partial graphs. The general question is then: *When* can an operation on relations over possible worlds be induced from some operation at the level of priority graphs? We merely state a conjecture:

4.4.15. Conjecture. *All definitions in PDL-format which preserve quasi-linear order on possible worlds are induced by definable operations on partially ordered priority graphs.*

4.5 An alternative format: 'Priority Product Update'

By now, we have looked at several operations that change given binary relations on possible worlds, or on objects in general. As observed at the beginning of this chapter, these relations can stand for many different things, from plausibility ordering in belief revision (cf. [Rot06]) to relative preference - and hence, techniques developed for one interpretation can often be used just as well for another. In particular, as we have noted, [Ben07a] used preference change as studied in our Chapter 2 for modeling various policies for belief revision. To achieve greater generality here, we have employed PDL-style definitions in the preceding Section 4.4 for relation transformers. In this section, we briefly consider an alternative approach from the recent literature on belief revision, which uses 'event models' as in dynamic-epistemic logic. Our discussion in Chapter 2 has introduced what event models $\mathcal{E}$ are, and the 'product update' $\mathcal{M} \times \mathcal{E}$ over given epistemic models $\mathcal{M}$ that is associated with them.

For a start, consider the following simple example, taken from [BS08]. The transformer $\Uparrow A$ may be naturally associated with an event model with two public announcements, or better in this setting: two 'signals' $!A$ and $!\neg A$, as shown in Figure 4.8, where the signal $!A$ is more plausible, or 'better', than $!\neg A$.

This describes a situation where we are not sure that A holds, but we do think that it is much more plausible than $\neg A$. How do we update with this event model? We need to define the new plausibility order on the pairs $\langle$*old world, new*

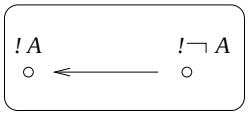

Figure 4.8: $\mathcal{E} = (!\neg A \leq !A)$

event$\rangle$ in $\mathcal{M} \times \mathcal{E}$. The general answer is the following rule of 'priority product update'. One might say that it works 'anti-lexicographically', giving priority to the event order (i.e. the last observation made), and world order only when the event order is indifferent.

4.5.1. Definition. (priority update). In product models $\mathcal{M} \times \mathcal{E}$, the plausibility relation on pairs (s, σ) is as follows:

$$(s, \sigma) \leq (s', \sigma') \quad \text{iff} \quad \sigma < \sigma' \quad \text{or} \quad \sigma \sim \sigma', s \leq s'.$$

[BS08] shows how this stipulation generalizes the examples in [Ben07a], while also dealing with a large variety of multi-agent belief revision scenarios. Moreover, the authors provide a dynamic doxastic language for which they prove completeness. Interestingly, one key element in their analysis is the use of simple modalities $\langle bett \rangle$ for the plausibility order, as well as an existential modality E, both as in our Chapter 2 ([BL07]) – as they in fact point out.[13]

Now the key innovation in this approach is the following shift. Instead of modeling different belief revision policies by different definitions, as we have done, there is just *one single update rule* which works in all circumstances. All further information about how to re-order the worlds more specifically has to be contained in the 'input signal', viz. the event model $\mathcal{E}$ with signals and priority ordering. One benefit of this approach is that one reduction axiom suffices for the basic modalities, instead of the different ones we gave for different policies.[14] This intriguing move shifts the generality from formats of definition for relation transformers to an account of the relevant event models. Moreover, event models have some formal analogies with our earlier priority graphs (for more discussion on this, we refer to [Gir08].).

There are some obvious questions about the relation between this event model format and our earlier PDL-style definitions. While things are not totally clear, and there may be a non-inclusion both ways, we can at least notice a few obvious facts. For the purpose of comparison, we stick with quasi-linear orders.

[13] Priority update also has the flavor of the above priority graph merge, and [Ben07b] gives a generalized formulation which also works for pre-orders.

[14] Of course, the latter still provide more concrete information about specific belief changes. And also, the precise status of the event models in this approach is a bit unclear, since they will often be no longer about real events from the original DEL motivation, but abstract signal combinations designed to encode revision policies.

4.5.2. FACT. Every PDL base definition defines a relation transformer whose action can also be defined as taking products $\mathcal{M} \times \mathcal{E}$ with some suitable event model $\mathcal{E}$ using Priority Product Update.

Proof. Here is a syntactic procedure for turning PDL base definitions into an equivalent event model. Since every such definition can be written in normal form by Fact 4.4.4, it is a finite union of relations between two possible worlds. The procedure goes as follows:

Step 1 Using standard propositional equivalences, rewrite the test conditions in the definition to become disjoint 'state descriptions' from some finite partition of the set of all worlds. We will write $?SD$ when referring to these. The definition then becomes a union of clauses $?SD; R; ?SD'$ and $?SD; \top; ?SD'$.

Step 2 Take the state descriptions as signals in an event model, and put an indifference relation between SD, SD' when the PDL definition has a clause $?SD; R; ?SD'$. Put a directed link from SD to SD' when the PDL definition has a clause $?SD; \top; ?SD'$.

In this way, information about the operation σ is moved into the event model $\mathcal{E}_\sigma$. Now it is easy to see that the following equivalence holds:

For any PDL definable operation σ, $\sigma(R^{\mathcal{M}}) = \leq^{BS}_{\mathcal{M}\times\mathcal{E}_\sigma}$.

The reason is that the indifference clause lets the old model decide, whereas the directed clause imposes new relations as required in the case of a clause $?SD; \top; ?SD'$.[15] □

Conversely, when is a priority update given by some event model $\mathcal{E}$ definable in our PDL format? Again, we just look at a special case, where the set of worlds does not change. The preconditions of the events in $\mathcal{E}$ then form a partition, and also, each event has a unique precondition.[16] In this setting, we also have a converse reduction. Here is the procedure:

Step 1 Let the event preconditions form a partition which is in one-to-one correspondence with the events themselves. This generates a PDL definition of the relation transformer where we test for the preconditions.

Step 2 Put $?SD; R; ?SD'$ when SD is indifferent with SD' in the event model.

[15] Note that this event model just copies the original model $\mathcal{M}$: unlike in general product update, no duplications occur of worlds via events, since all SD are mutually exclusive and together exhaustive.

[16] This is a bit like the scenario for 'protocols' in [BGK06].

Step 3 Put $?SD; \top; ?SD'$ when SD' is preferred to SD in the event model.

It is easy to see, using the analogy with the converse procedure above, that this proves the following

4.5.3. FACT. Priority product update on partition event models induces relation changes which are definable in the basic PDL format.

More can be said here. For instance, if the order on $\mathcal{E}$ is quasi-linear, then only PDL definitions are relevant which preserve quasi-linearity. Indeed, many PDL-definitions produce only pre-orders, and hence priority product update would need to be generalized (cf. [Ben07b]). And on top of that, definitions with iterated occurrences of the input relation R, while perfectly fine from the viewpoint of finding reduction axioms (cf. [BL07]) have no obvious product update counterparts. Conversely, if we were to write PDL-definitions for whole event models, we would need to generalize the [BL07] relation transformers to a setting where the operation does not just change the relation among the existing objects, but also may create new objects and drop old ones.

Our analysis suggests that PDL operations and priority product update have related but still somewhat different intuitions about achieving generality, and their connection is not yet totally clear.

4.6 Comparing logical languages

So far we have compared the proposals from the previous two chapters mainly from a *semantical* point of view. But 'logic' only arises when we also introduce *formal languages* to talk about these semantic structures. This section is meant to draw some comparisons on the formalisms that have been used. We will use some tables to summarize the main features.

The language in Chapter 2 is a modal language over combined epistemic - betterness models, with K as its standard epistemic operator, and a less standard modal operator $[bett]$ for describing 'local preferences'. Following [BOR06], a further hybrid universal modality U was added, to better express, in combination with $[bett]$, various notions of preference between propositions. This is the static part of the complete language. It was used to describe standard modal models $(W, \sim, \preceq, V)$, where, as usual, W is a set of possible worlds, $\sim$ an equivalence relation for knowledge, $\preceq$ the 'at least as good as' relation, and V the atomic valuation function.

In addition, we high-lighted dynamic changes of models in Chapter 2. To do so, two dynamic operators were included in the language: viz. modalities for public announcements $[A!]$ and suggestions $[\sharp A]$. For instance, the formula $[\sharp A]\varphi$ expresses that 'after a suggestion A, φ holds'. Typically, we had a complete set of

reduction axioms to speak about the changes before and after a dynamic action. This format for relation change was even extended to PDL-definable changes. In a diagram, here are all the mentioned ingredients:

Static	`modal`	`new operator`	`hybrid`	`preference defined`
	$p \mid \neg \mid \wedge \mid K$	$[bett]$	U	$Pref^{\forall\exists}$,$Pref^{\forall\forall}$, etc.
Dynamic	$A!$	$\sharp A$		

The language in Chapter 3 was not modal, but rather a fragment of a first-order doxastic language. As shown again in the table below, it has two levels. The 'reduced language' consists of propositional formulas, and expressions of preference over constants: e.g. $Pref(d_i, d_j)$ said that 'd_i is preferable over d_j'. In the extended language, we then added first-order quantifiers, predicates, and variables, allowing us to talk about priorities explicitly. Moreover, we introduced an operator for agents' beliefs. Here the intended semantic models are first-order doxastic structures $\langle W, D, R, \{\preceq_w\}_{w\in W}, V\rangle$, where W is a set of worlds, D a set of distinguished constant objects, and R a euclidean and serial accessibility relation on W. For each w, $\preceq_w$ is a quasi-linear order on D, which is the same throughout each euclidean equivalence class, and V is again an atomic valuation function.

Similarly to Chapter 2, Chapter 3 also explored dynamics. Beliefs get changed through change in plausibility structure, just as in [Ben07a], with the revision operator $\Uparrow A$ as an example. Typically, belief revision leads to a preference change, since we defined preference in terms of beliefs. But also, the priority sequence can be changed directly, leading to a second kind of preference change. For this purpose, we introduced modalities for the earlier-mentioned four operations: $[^+A]$ for adding A to the right, $[A^+]$ for adding A to the left, $[-]$ for dropping the last element of a priority sequence, and $[i \leftrightarrow i{+}1]$ for interchanging the i-th and $i{+}1$-th elements. Again, a complete set of reduction axioms has been given for these operators. This time, the relevant Table is:

Static	`reduced language`	`extended language`	`priorities`
	$p \mid \neg \mid \wedge \mid Pref \mid B$	$P(d_i) \mid x \mid \forall$	$P(x) \gg Q(x)$
Dynamic	$\Uparrow A$	$[^+A], [A^+], [-], [i \leftrightarrow i+1]$	

This table may actually suggest a linguistic 'gap'. We have used priority sequences extensively in Chapter 3, but they have never become first-class citizens in the language. If one wanted to develop our ideas more radically, one might use a language for talking about the priorities themselves, their order, principles for reasoning about or with them, or even for changing them.

While this looks like a stark omission, things are actually much brighter. In fact, many of the previous semantic observations are already valid principles of

a *calculus of priority sequences*, or even of priority graphs. For instance, the following equivalences were shown valid (Fact 4.2.7, Fact 4.4.9):

1. $\preceq_{\mathscr{P} \uplus \mathscr{P}'} = \preceq_{\mathscr{P}} \cap \preceq_{\mathscr{P}'}$.
2. $\preceq_{A;\mathscr{P}} = \Uparrow A(\preceq_{\mathscr{P}})$.

But is is clear that these are algebraic laws of some kind, provided we introduce the right notation. We see such a calculus as a natural follow-up to Chapter 3. Actually, [Gir08] has started a related investigation on 'agenda change' based on [ARS02]. We refer to Chapter 3 and [Gir08] for further details.

What about comparisons between the logical formalisms employed in Chapters 2 and 3? The difference between modal and first-order is not crucial here, as is well-known from modal correspondence theory (cf. [Ben99], [ABN98] and [BB07]). In Chapter 2, our basic concern is the betterness relation on the possible world level, and the modal language describes its 'local properties' at individual worlds. If one wants to make more global assertions, e.g. about propositional preference, unrestricted quantifiers are needed, and the universal modality U was a half-way station to full first-order logic here. But one can also use the first-order language of Chapter 3 in the end, as well as various fragments of it, modal or non-modal. Of course, things get more complex when we consider dynamic model-changing operators, as we would have to compare dynamic modal and first-order languages. Finally, when a language is added which talks about priorities, i.e. about propositions as objects, then we can view this either as a mild form of second-order logic, or as a *two-sorted first-order language* over our two-level structured models of Section 4.1. And the latter language will again have obvious modal fragments. Thus we conclude that, appearances notwithstanding, the languages used in Chapters 2 and 3 are very close.

Finally, the more interesting question is maybe this. Given the semantic connections between the structures employed in Chapters 2 and 3, and the connections between their languages, can we also find explicit *reductions between the logics* that we have proposed in these two separate investigations? We think one can, but we leave this matter to future investigation. Instead, we conclude this chapter with two further topics. One is the question how the entanglement of preference and belief, which was so central in Chapter 3, should play a role in the modal languages of Chapter 2. The other is the issue how one could merge all ideas from Chapters 2 and 3 into one logical system that might have the power to address preference in much greater generality.

4.7 Preference meets belief

We have seen just now in Section 4.6 that preference is not just a matter of pure 'betterness'. In addition, it involved epistemic operators K of knowledge in

Chapter 2 and doxastic operators B of belief in Chapter 3. Understanding this entanglement of preference with knowledge and belief is of importance, especially when we study how agents make choices under uncertainty. This section compares the perspectives of the previous two chapters in this regard.

Preference is usually defined explicitly in terms of beliefs when we only have incomplete information. Here is a brief review of how this worked with the main notions of Chapter 3:

(i) We distinguished preference over *objects* and preference over *propositions*. First, preference over objects was defined by beliefs on whether objects have certain properties. We then applied the same method to preference over propositions.

(ii) For ease of reading, we repeat some basic definitions here:

Given a priority sequence $\mathscr{P}$ of the following form

$$P_1(x) \gg P_2(x) \gg \cdots \gg P_n(x) \quad (n \in \mathbb{N}),$$

where each of the $P_m(x)$ is a formula from the language, all with one common variable x, we define preference over objects as follows:

$$Pref(d_1, d_2) ::= \exists P' \in \mathscr{P}(\forall P < P'(BPd_1 \leftrightarrow BPd_2) \wedge (BP'd_1 \wedge \neg BP'd_2)). \qquad (Pref\text{-}obje)$$

(iii) Preference over propositions was defined similarly. Given a propositional priority sequence of length n of the following form

$$\varphi_1(x) \gg \varphi_2(x) \gg \cdots \gg \varphi_n(x) \quad (n \in \mathbb{N}),$$

where each $\varphi_m(x)$ is a propositional formula with an additional propositional variable, we define preference over propositions ψ and θ as follows:

$$Pref(\psi, \theta) \quad \text{iff} \quad \text{for some } i \ (B(\varphi_1(\psi) \leftrightarrow B(\varphi_1(\theta)) \wedge \cdots \wedge (B(\varphi_{i-1}(\psi)) \leftrightarrow B(\varphi_{i-1}(\theta))) \wedge (B(\varphi_i(\psi) \wedge \neg B(\varphi_i(\theta))). \qquad (Pref\text{-}prop)$$

(iv) We took the line that preference is a state of mind (i.e. it is subjective, and subject to introspection) and therefore, one prefers one alternative over another if and only if one believes one does. So typically, $Pref(d_1, d_2) \leftrightarrow BPref(d_1, d_2)$ was an axiom of our preference logic, which therefore includes positive introspection.

In contrast, beliefs are not a part of the language considered in Chapter 2, which only has the universal modality U, knowledge operator K and the betterness modality $[bett]$. This vocabulary allows us to express many notions of preference over propositions, depending on different combinations of quantifiers ('liftings'). Since the betterness operator is based on an objective relation in the model, the related preference modality will in general not be subjective. However, we can prefix it with epistemic operators, and achieve attitude-dependence after all. Thus, we discussed the connection between preference and knowledge, asking, e.g. whether epistemized preference validates positive introspection. Also, we have looked at the particular notion of 'regret', interpreted as 'agent a knows that p but she prefers that $\neg p$', which is only possible by combining objective and subjective aspects.

To make the preceding two approaches more comparable, we will first add beliefs to Chapter 2 and develop the system a bit further. After that, we will draw a comparison between Chapter 2 and Chapter 3, and their different notions of preference. First, we briefly review some important technical points from Chapter 2 that will return later.

(i) The language can express preferences over propositions, viewed as sets of possible worlds. We defined the central notion of $Pref^{\forall\exists}$ as follows:

$$Pref^{\forall\exists}(\varphi,\psi) ::= U(\psi \to \langle bett\rangle\varphi). \quad (Ubett)$$

(ii) A new reduction axiom for the operator $\langle bett\rangle$ was proposed to model changes in preference relations, with the running example of a 'suggestion' A ($\sharp A$). Given the reduction axioms below, this also determines the complete logic of changing preferences between propositions.

1. $\langle \sharp A\rangle\langle bett\rangle\varphi \leftrightarrow (\neg A \wedge \langle bett\rangle\langle \sharp A\rangle\varphi) \vee (\langle bett\rangle(A \wedge \langle \sharp A\rangle\varphi))$.
2. $\langle A!\rangle\langle K\rangle\varphi \leftrightarrow (A \wedge \langle K\rangle\langle A!\rangle\varphi)$.
3. $\langle A!\rangle E\varphi \leftrightarrow (A \wedge E(\langle A!\rangle\varphi \vee \langle \neg A!\rangle\varphi))$.

(iii) Since $Pref^{\forall\exists}$ is defined by the operators U and $\langle bett\rangle$, once we have reduction axioms for these two operators separately, we immediately get one for $Pref^{\forall\exists}$. The precise calculation went as follows:

$$\begin{aligned}
&\langle \sharp A\rangle Pref^{\forall\exists}(\varphi,\psi) \leftrightarrow \langle \sharp A\rangle U(\psi \to \langle bett\rangle\varphi)\\
&\leftrightarrow U(\langle \sharp A\rangle(\psi \to \langle bett\rangle\varphi))\\
&\leftrightarrow U(\langle \sharp A\rangle\psi \to \langle \sharp A\rangle\langle bett\rangle\varphi)\\
&\leftrightarrow U(\langle \sharp A\rangle\psi \to (\neg A \wedge \langle bett\rangle\langle \sharp A\rangle\varphi) \vee (\langle bett\rangle(A \wedge \langle \sharp A\rangle\varphi)))\\
&\leftrightarrow U(\langle \sharp A\rangle\psi \wedge \neg A \to \langle bett\rangle\langle \sharp A\rangle\varphi) \wedge U(\langle \sharp A\rangle\psi \wedge A \to \langle bett\rangle(\langle \sharp A\rangle\varphi \wedge A))\\
&\leftrightarrow Pref^{\forall\exists}(\langle \sharp A\rangle\varphi, \langle \sharp A\rangle\psi) \wedge Pref^{\forall\exists}((\langle \sharp A\rangle\varphi \wedge A), (\langle \sharp A\rangle\psi \wedge A)).
\end{aligned}$$

Preference and knowledge

As we saw in the above, the preference in Definitions (*Pref-obje*) and (*Pref-prop*) is based completely on beliefs, and hence Chapter 3 took a subjective stance. But intuitively, the preference defined in (*Ubett*) is more *objective*. In comparing this, for the moment, we ignore the difference between objects and possible worlds, but we will come back to it at the end of this section. Intuitively, (*Ubett*) says the following:

> for any ψ-world in the model, there exists a world which is at least as good as that world, where φ is true.

This can be pictured as follows:

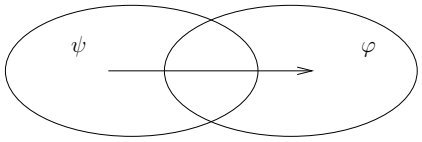

Figure 4.9: Preference defined by U and betterness relations

Essentially this is a comparison between ψ-worlds and φ-worlds in the model, with no subjective attitude involved yet. But even in the setting of Chapter 2, we can create connections between preference, knowledge, and beliefs - as we are going to show now.

Consider the following situation. Instead of picking *any* ψ-world in the model (as the universal modality U does), we now only look at those ψ-worlds that are *epistemically accessible* to agent i. This suggests the following intuition:

> For any ψ-world that is *epistemically accessible* to agent i in the model, there exists a world which is as good as that world, where φ is true.

This can be pictured in the following manner:

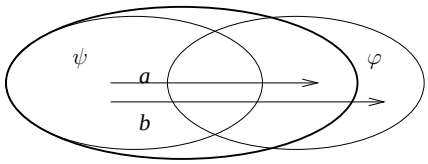

Figure 4.10: Preference defined by K and betterness relations

The part inside the black circle stands for the epistemically accessible worlds. The other two circles in the picture stand for the set of ψ-worlds, and the set of φ-worlds, respectively. So only some of the φ-worlds are epistemically accessible. The betterness relation has two possible cases: either it consists of a-arrows,

which means that the better φ-world is itself in the accessible part of the model, or it also consists of b-arrows, which means that the better φ-world need *not* be in the accessible part of the model.

We write the above explanation in the formal language as:

$$Pref^{\forall\exists}(\varphi,\psi) ::= K_i(\psi \to \langle bett\rangle\varphi). \quad (Kbett)$$

Comparing the definitions $(Kbett)$ and $(Ubett)$, we have simply replaced U with K_i. In fact, looking back at Chapter 2, this is a straightforward step to take, since we had a knowledge operator in the language. The models proposed in Chapter 2 can be used directly for this purpose, and likewise, their complete logics.

Next, regarding dynamics, preference change is now triggered by changes in both epistemic and betterness relations. And we can obtain the right reduction axiom for epistemic preference in the same manner as for the universal modality, by a calculation from the reduction axioms for K_i and $\langle bett\rangle$. It is easy to spell out the details.

Introducing beliefs

In many situations, however, we do not have solid knowledge, but only beliefs. Still we want to compare situations in terms of betterness. In other words, we would like to say things like this:

> for any ψ-world that is most plausible to agent i in the model, there exists a world which is as good as that world, where φ is true.

This requires introducing beliefs – formally at first:

$$Pref^{\forall\exists}(\varphi,\psi) ::= B_i(\psi \to \langle bett\rangle\varphi). \quad (Bbett)$$

Figure 4.11 illustrates what we have in mind now:

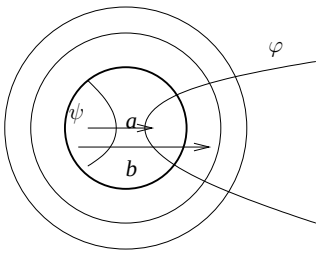

Figure 4.11: Preference defined by B and betterness relations

In the picture, the worlds lie ordered according to their plausibility, as in Lewis' spheres for conditional logic. The part inside the black circle depicts the most plausible worlds. We consider the ψ-worlds in this area, and again

distinguish two sorts of preference relations: relations of 'type a' stay inside the most plausible region, relations of 'type b' go outwards to the less plausible, or even implausible region of the model.

To interpret beliefs more formally, we define these models as follows:

4.7.1. Definition. A *belief preference model* is a tuple $\mathcal{M} = (W, \leq, \preceq, V)$, with W a set of possible worlds, $\leq$ a doxastic relation of 'at least as plausible as', and $\preceq$ our earlier relation of 'at least as good as', with V again a valuation for proposition letters.[17]

The truth conditions for the absolute belief operator B and more general conditional beliefs $B^{\psi}\varphi$ are defined as follows:

> $\mathcal{M}, s \models B\varphi$ iff $\mathcal{M}, t \models \varphi$ for all worlds t which are minimal for the ordering $\lambda xy.\ \leq_s xy$.

> $\mathcal{M}, s \models B^{\psi}\varphi$ iff $\mathcal{M}, t \models \varphi$ for all worlds t which are minimal for $\lambda xy.\ \leq_s xy$ in the set $\{u \mid \mathcal{M}, u \models \psi\}$.

Here the truth condition for the unary operator B is essentially the same as in $KD45$-models, with the 'accessible worlds' of the latter system being the most plausible ones. But here we can compare less plausible worlds, too, and this is crucial to understanding conditional belief.

4.7.2. Remark. Given the notion of conditional belief, there is actually an alternative formulation for our formulation of belief-based preference. The above version (*Bbett*) looks at all normal or optimal worlds in the model, and then compares φ-worlds to ψ-worlds there in terms of betterness. The other option would be this: take the preference for ψ over φ itself as a *conditional belief*, using the following formula

$$B^{\psi}\langle bett\rangle\varphi \qquad (Bbett').$$

As is well-known, this is not equivalent to (*Bbett*), and it might be another candidate for belief-based preference. Personally, we think that preference should not involve the conditional scenario of 'having received the information that ψ'. However, both definitions can be treated in the logic we have proposed, and both are amenable to the style of dynamic analysis that we will consider next.

Again, we can now model changes in two ways, through changes in the plausibility relation and through changes in the betterness relation of the model. The

[17] [BS06b] and [BS08] also uses the 'as plausible as' relation to interpret the notion of *safe beliefs* which hold in all worlds that are at least as plausible as the current one. This notion is like our universal betterness modality, but then of course for belief rather than preference.

DEL methodology still applies here, since the two cases are formally very similar. Accordingly, [Ben07a] proposed valid reduction axioms for two sorts of belief change, so-called *radical* revision and *conservative* revision, where the former involves our earlier relation transformer $\Uparrow A$. These technical results can be used here directly. For instance, the reduction axiom for beliefs after radical revision is:

$$[\Uparrow A]B\varphi \leftrightarrow (EA \wedge B([\Uparrow A]\varphi|A)) \vee B[\Uparrow A]\varphi.$$

For a complete system, we also need a reduction axiom for conditional belief. The details are in [Ben07a].

In the same line, once we have reduction axioms for the belief and betterness operators, we can calculate what should be the reduction axiom for defined preference over propositions, e.g., the $Pref^{\forall\exists}$ notion in ($Bbett$). We then obtain a complete logic for dynamical change of belief-based preference.

Definitions ($Kbett$) and ($Bbett$) share a common feature: an arrow to a better φ-world can lead *outside of the accessible or most plausible part* of the model, witness the earlier arrows of type b. The intuition behind this phenomenon is clear, and reasonable in many cases. It may well be that there exists better worlds, which the agent does not view as epistemically possible, or most plausible. But *if* we want to have the two base relations entangled more intimately, we might want to just look at better alternatives inside the relevant epistemic or doxastic zone. Such considerations are found in the study of normative reasoning in [LTW03] where a normality relation and a preference relation live in one model. Likewise, [BRG07] discuss the 'normality sense' of ceteris paribus preference, restricting preference relations to just the normal worlds for the agents. In what follows, we will explore this more intimate interaction of the two base relations a bit more.

Merging relative plausibility and betterness

Now we require that the better worlds relevant to preference stay inside the most plausible part of the model. Intuitively, this means that we are 'informational realists' in our desires. To express this, we need a merge of the two relations, viz. their intersection. Here is how:

4.7.3. Definition. A *merged preference model* is a tuple $\mathcal{M} = (W, \leq, \preceq, \leq \cap \preceq, V)$, with W a set of possible worlds with doxastic and betterness relations, but also $\leq \cap \preceq$ as the intersection of the relations 'at least as plausible as' and 'at least as good as', with V again a valuation for proposition letters.

The original language had separate modal operators B and $[bett]$, but now we extend it with a new modality H. The formula $H\varphi$ is interpreted as 'it is hopeful that φ'. The truth condition for such formulas is as follows:

> $\mathcal{M}, s \models H\varphi$ iff for all t with both $s \leq t$ and $s \preceq t$, it holds that $\mathcal{M}, t \models \varphi$.

With this new language over these new models, we can define one more natural notion of preference over propositions, which is actually much closer to Definition (*Pref-prop*) from Chapter 3:

$$Pref^{\forall\exists}(\varphi, \psi) ::= B(\psi \to \langle H\rangle\varphi). \quad (BH)$$

In words, this says that:

> For any most plausible ψ-world in the model, there exists a world which is *as good as* this world, and at the same time, *as plausible as* this world, where φ is true.

Obviously, we can now talk about preferences restricted to the most plausible part of the model. In terms of Figure 4.11, only arrows of 'type a' remain.

Actually, this same move would apply to Definition ($Kbett$) as well. Requiring that the better worlds stay inside the accessible worlds, we would have:

$$Pref^{\forall\exists}(\varphi, \psi) ::= K(\psi \to \langle \sim \cap \preceq\rangle\varphi) \quad (Kbett').$$

This means that we keep only the a-arrows in Figure 4.10.

Definition (BH) gives us a subjective notion of preference, as we consider only the most plausible part of the model. As we said, it is getting closer to Definition (*Pref-prop*). More precisely, Remark 3.7.10 showed that the $\forall\exists$ version of the preference defined by (*Pref-prop*) is equivalent to the following

$$B(\psi \to \langle bett\rangle\varphi).$$

Since then $[bett]\varphi ::= \underline{Pref}(\varphi, \top)$ and $\underline{Pref}$ is defined in terms of beliefs, the betterness relation automatically stays within the plausible part of the models. Hence, $\forall\exists$-(*Pref-prop*) is actually equivalent to Definition (BH).

Now let us quickly look at the expressive power of the modal language with a new operator H. Can the notion of preference in (BH) be defined in the original language with modal operators B and $[bett]$ only? In other words, can iterations of separate doxastic and betterness modalities achieve the same effect as intersection? As we know from general modal logic, this is very unlikely, since intersection modalities are not invariant under bisimulation (cf. [BRV01]). Indeed, the answer is negative:

4.7.4. Fact. $B(\psi \to \langle H\rangle\varphi)$ $(*)$ is not definable in the standard bimodal language with modal operators B and $[bett]$.

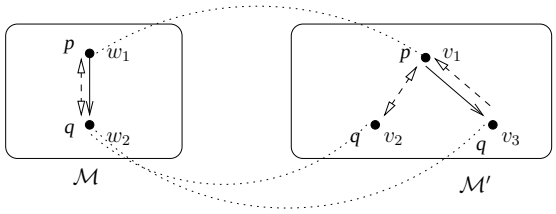

Figure 4.12: Bisimilar models

Proof. Suppose $(*)$ were definable. Then there would be a formula φ in the language without H such that $\varphi \leftrightarrow (*)$ holds in every model. Now consider the two models in Figure 4.12.

The betterness relation $\preceq$ is pictured by solid lines with arrows, and the plausibility relation $\leq$ by dashed lines with arrows. The evaluation of the proposition letters p and q can be read off from the picture. It is easy to see that these two models are bisimilar with respect to both betterness and relative plausibility, with the bisimulation indicated by the dotted lines.

Now, we have $\mathcal{M}, w_1 \models B(p \to \langle H\rangle q)$, since the p-world w_1 can see a world w_2 which is both better and plausible where q is true. Then we should get $\mathcal{M}, w_1 \models \varphi$, since $\varphi \leftrightarrow (*)$. Because $\mathcal{M}$ and $\mathcal{M}'$ are bisimilar, we would then have $\mathcal{M}', v_1 \models \varphi$. So we should also have $\mathcal{M}', v_1 \models B(p \to \langle H\rangle q)$. But instead, we have $\mathcal{M}', v_1 \not\models B(p \to \langle H\rangle q)$, because the p-world v_1 can see v_2 which is plausible but not better, and v_3 which is better but not plausible. So there is no world which is both better and plausible, while satisfying q. This is a contradiction. □

This argument shows that the new language indeed has richer expressive power. While this is good by itself, it does raise the issue whether our earlier methods still work.

In particular, we consider possible dynamic changes to the merged relation. We will only look at our three characteristic actions: radical revision $\Uparrow A$ that changes the plausibility relations, suggestion $\sharp A$ that changes the betterness relations, and the standard public announcement $A!$ that changes the domain of worlds. As it happens, the DEL-method of reduction axioms still applies:

4.7.5. Theorem. *The following equivalences are valid:*

1. $\langle \sharp A\rangle\langle H\rangle\varphi \leftrightarrow (A \wedge \langle H\rangle(A \wedge \langle \sharp A\rangle\varphi)) \vee (\neg A \wedge \langle H\rangle\langle \sharp A\rangle\varphi)$.
2. $\langle \Uparrow A\rangle\langle H\rangle\varphi \leftrightarrow (A \wedge \langle H\rangle(A \wedge \langle \Uparrow A\rangle\varphi)) \vee (\neg A \wedge \langle H\rangle(\neg A \wedge \langle \Uparrow A\rangle\varphi)) \vee (\neg A \wedge \langle bett\rangle(A \wedge \langle \Uparrow A\rangle\varphi))$.
3. $\langle A!\rangle\langle H\rangle\varphi \leftrightarrow A \wedge \langle H\rangle\langle A!\rangle\varphi$.

Proof. We only explain the most interesting Axiom 2 as an illustration. Assume that $\langle \Uparrow A\rangle\langle H\rangle\varphi$. Recall that radical revision $\langle \Uparrow A\rangle$ only changes the plausibility

relation, leaving the preference relation intact. The new plausibility relation was written as follows:

$$(?A; R; ?A) \cup (?\neg A; R; ?\neg A) \cup (?\neg A; \top; ?A)$$

Seen from the initial model, we can therefore distinguish three cases, and these are just the three disjuncts on the right-hand side. Note that for the last one we only need to insert the old preference relation $\langle bett \rangle$, since the plausibility relation $(?\neg A; \top; ?A)$ is new. □

As for axiomatizing the complete logic of the new modality H, there are various techniques. For instance, one can introduce 'nominals' from hybrid logic, as is done in axiomatizing modal logics with intersection modalities (cf. [Bla93], [Kat07]). The preceding observation then shows that the complete dynamic logic can be obtained by adding just these three reduction axioms. Thus, the DEL-methodology also works in this extended setting.

For the reader's convenience, we tabulate the many different definitions of preference that we have seen so far, and the implicational relationships between them in the following:

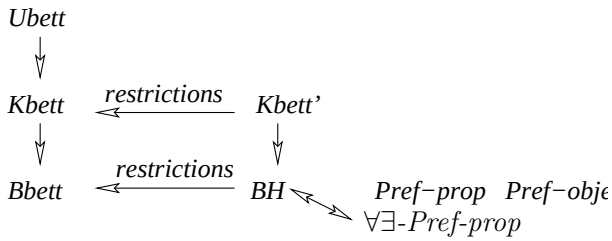

Figure 4.13: Different definitions of preference

We started with the pure betterness Definition ($Ubett$) from Chapter 2. On the same pattern, we 'epistemized' the universal modality, and proposed Definitions ($Kbett$) and ($Bbett$). The relation between these is as follows: ($Ubett$) implies ($Kbett$), ($Kbett$) implies ($Bbett$), but no reverse implication holds. Then, we set restrictions to make the betterness comparisons stay inside the accessible or most plausible region of the model, obtaining new Definitions ($Kbett'$) and (BH). Their relations to the previous notions are as indicated. Finally, as for connections with Chapter 3, defining preference as in Definitions ($Pref\text{-}prop$) makes them comparable to Definition (BH). This is true technically, but also intuitively, since all these notions are subjective. In particular, we have shown that Definition (BH) is equivalent to a $\forall\exists$ lifted version of Definition ($Pref\text{-}prop$).

4.8 Combining preference over objects and preference over possible worlds

The preceding section concludes our comparisons between the two levels for defining preference in Chapters 2 and 3. But there is one more viewpoint which we want to mention briefly. Instead of taking these as different approaches that need to be contrasted and compared, one could equally well say that they bring out equally natural aspects of preference which need to be *merged*. And indeed, it is easy to do so. In this final section, we merely show how this may be done consistently, and what further questions would arise.

First, consider the following difference in 'spirit'. Priority sequences naturally fit with preferences between *objects*, while the propositional modal languages used so fit better with preferences between *worlds*. As we have said, technically, this does not make much difference. Theorem 3.7.12 even shows an equivalence between preference over objects and preference over propositional variables. Nevertheless, in real life we often compare objects and situations at the same time, neither exist exclusively. And these express different things. Consider the following example:

4.8.1. Example. Alice prefers living in Amsterdam over living in Beijing. In Amsterdam there are two houses d_1 and d_2 for her to choose from, and she prefers d_1 over d_2. In Beijing, she prefers house d_3 over d_4.

A more abstract example of the distinction would be this. In some worlds, we may have many preferences between objects of desire, while in others, we have none at all. In some philosophies and religions, we would prefer the worlds where we have few object preferences (or better: none) to those where we have many. Moreover, things get even more complex when we bring in relative plausibility and beliefs. E.g. Alice may prefer living in Amsterdam, while still thinking it more plausible that she will end up domiciled in Beijing - or vice versa.

To talk about such examples, we need to combine preference over objects and preference over possible worlds in one semantic structure. The proper vehicle for this would join our earlier languages into one *doxastic preferential predicate language* defined as follows:

4.8.2. Definition. Object-denoting terms t are variables $x_1, x_2, \ldots$ and constants $d_1, d_2, \ldots$, $P_1, P_2, \ldots$ are predicates over objects. The *language* is defined in the following syntax format:

$$\varphi ::= Pt_1, \ldots, t_n \mid \neg\varphi \mid \varphi \wedge \psi \mid \underline{Pref}(t_i, t_j) \mid B\varphi \mid [pref]\varphi$$

This is only a small part of a complete doxastic preferential predicate logic, but it is already adequate for many purposes. Of course, one can extend this language with quantifiers $\exists x\varphi$ in a straightforward manner.

Semantic models appropriate to this language may be defined as follows:

4.8.3. DEFINITION. A *preferential doxastic predicate model* is a tuple $\mathcal{M} = (W, \leq, \preceq, \{D_w \mid w \in W\}, \{\preceq_w \mid w \in W\}, V)$, with S a set of possible worlds, $\leq$ and $\preceq$ a plausibility relation and a betterness relation over these. Next, D_w is the domain of objects for each possible world $w \in W$, with $\preceq_w$ a distinguished relation 'at least as good as' over objects in these domains. Finally, V is a valuation or interpretation function for the constants and predicate atoms of the language.

This means that we have preferences between worlds, and inside the possible worlds, we have preference over objects. Moreover, worlds are also ordered according to their plausibility. In this way, preference lives at different levels of the semantic models. The operators of the language can now be interpreted as usual. In particular, if we were to add quantifiers, these would range over the local domains D_w. Notoriously, there are difficult issues in the semantics and proof theory of modal predicate logic (cf. [FM98], [HC96], and [BG07]), but even so, a framework like this seems needed to discuss more subtle points of preference.

For instance, in a language like this we can now state assertions like:

(a) agents believe that objects with property P are always better than objects with property Q,

(b) agents prefer situations where object d is not preferred to object e,

(c) agents prefer situations where they do not know if P to situations where they do know if P.

This would seem to be the expressive power needed to do justice to discussions in the philosophical literature like, [Han90b], [Åqv94] and [Han01a]. Also, this setting gets closer to complex scenarios like the 'deontics of being informed' investigated in [PPC06].

But we can go even further. One might even compare objects across possible worlds, as in Russell's famous sentence "I thought your yaught was longer than it is". Alice might prefer her favorite house in Amsterdam when she lives there to her favorite house in Beijing when she does not live there: 'the grass is always greener on the other side'. And finally, we could also make the priorities of Chapter 3 into explicit elements of the semantics, letting the language speak about whether or not agents believe the priorities which determine their preferences over objects.

Clearly, this richer setting also raises issues of how to do dynamics. For instance, in the Example 4.8.1, both preference over objects and preference over possible worlds may change when new information comes in. We believe that the dynamic logics in our previous chapters can be generalized to deal with this, but there is hardly any work in this direction. For a recent *DEL*-style approach to

modal predicate logic, see [Koo07] on the dynamic semantics of changing assignments as well as worlds.[18]

Our conclusion is this. Merging the approaches in Chapters 2 and 3 is quite feasible, and comparisons at a semantic and modal propositional level reveal many analogies and compatible features. But doing this in full generality would require a modal predicate-logical framework, which seems feasible, but beyond the horizon of this study.

[18] Also relevant are earlier analysis in PDL terms by [EC92] and [Ben96].

Chapter 5

Diversity of Logical Agents in Games

In the preceding chapters, we have seen how agents can have quite diverse preferences, based partly on potentially highly different beliefs. Moreover, they may have different ways of changing these preferences and beliefs. But this is only the beginning of a much longer story of *agent diversity*. It is typical of rational agency that we are *not* all the same, and nevertheless, manage to coordinate activities and exchange information in successful ways. In this chapter, we will look for further clues for this diversity in the setting of concrete activities, viz. *games*. In particular, in games, even before we can get to players' beliefs and preferences, there is the simple basic mechanism of observation of moves that are played, and the information which comes to players because of this. This brings us to sources of diversity having to do with epistemic logic, which will be the main topic pursued here - though we will also provide more material on belief revision in the end as well.

5.1 Introduction: varieties of imperfection

Logical agents are usually taken to be epistemically perfect. But in reality, imperfections are inevitable. Even the most logical reasoners may have limited powers of observation of relevant events, generating uncertainty as time proceeds. In addition, agents can have processing bounds on their knowledge states, say, because of finite memory capacities. This chapter is an exploration of how different types of agents can be described in logical terms, and even co-exist inside the same logical system. Our motivating interest in undertaking this study concerns games with imperfect information, but our only technical results so far concern the introduction of imperfect agents into current logics for information update and belief revision. For a more extensive discussion of issues concerning diversity of agents, we refer to Chapter 6.

5.2 Imperfect information games and dynamic-epistemic logic

Dynamic-epistemic language Games in extensive form are trees $(S, \{R_a\}_{a \in A})$, consisting of nodes for successive states of play, with players' moves represented as binary transition relations between nodes. Imperfect information is encoded by equivalence relations $\sim_i$ between nodes that model uncertainties for player i. Nodes in these structures are naturally described in a combined *modal-epistemic language*. An action modality $[a]\varphi$ is true at a node x when φ holds after every successful execution of move a at x, and a knowledge modality $K_i\varphi$ is true at x when φ holds at every node $y \sim_i x$. As usual, we write $\langle a \rangle$, $\langle K \rangle$ for the existential duals of these modalities. Such a language can describe many common scenarios.

5.2.1. Example. In the following two-step game tree, player $\boldsymbol{E}$ does not know the initial move that was played by $\boldsymbol{A}$:

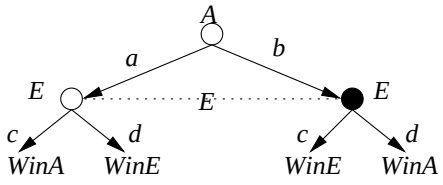

Figure 5.1: Not knowing one's winning move

The modal formula $[a]\langle d \rangle Win_E \wedge [b]\langle c \rangle Win_E$ expresses the fact that $\boldsymbol{E}$ has a winning strategy in this game, and at the root, she knows both conjuncts. After $\boldsymbol{A}$ plays move b from the root, however, in the black intermediate node, $\boldsymbol{E}$ knows merely 'de dicto' that playing either c or d is a winning move, as is expressed by the joint modal-epistemic formula $K_E(\langle c \rangle Win_E \vee \langle d \rangle Win_E)$. But she does not know 'de re' of any specific move that it guarantees a win: $\neg K_E\langle c \rangle Win_E \wedge \neg K_E\langle d \rangle Win_E$ also holds. In contrast, given the absence of dotted lines for $\boldsymbol{A}$, whatever is true at any stage of this game is known to $\boldsymbol{A}$. In particular, at the black intermediate node, $\boldsymbol{A}$ does know that c is a winning move for $\boldsymbol{E}$.

5.2.2. Remark. (temporal language). For some purposes, it is also useful to have *converse* relations $a^{\cup}$ for moves a, looking back up into the tree. In particular, these help describe play so far by mentioning the moves that have been played, while they also allow us to look back and say what could have happened if play had gone differently. Both are very natural things to say about the course of a game. This is a simple temporal logic variant of the basic modal-epistemic language. For a recent take up in this direction, we refer to [Yap06].

Strategies, plans, and programs A modal-epistemic language describes players' moves and what they know about their step-by-step effects in a game. Explicit information about agents' global behaviour can be formulated in a *dynamic-epistemic language*, which adds complex program expressions. A *strategy* for player i is a function from i's turns x in the game to possible moves at x, while we might think of a *plan* as any relation constraining these choices, though not always to a unique one. Such binary relations and functions can be described using the following expressions

(i) single moves a,

(ii) tests $(\varphi)?$ on the truth of some formula φ,

(iii) use of operations union $\cup$, relational composition ;, and iteration *.

In particular, these operations define the usual slightly more complex program constructs *IF THEN ELSE* and *WHILE DO*. As for test conditions, in this setting, it only makes sense to use φ which an agent *knows to be true or false*. Without loss of generality, we can assume that such conditions have the epistemic form $K_i\varphi$. The resulting programs are called 'knowledge programs' in [FHMV95]. [Ben01] proves that in finite imperfect information games, the following two notions from logic and game theory coincide:

(a) strategies that are defined by knowledge programs,

(b) *uniform strategies*, where players choose the same move at every two nodes which they cannot distinguish.

Valid laws of reasoning about agents and plans Universally valid principles of our language consist of the minimal modal or dynamic logic, and the epistemic logic matching the uncertainty relations – in our case, multi-S5. Logics like this were used in [Moo85] to study planning agents in AI. Of course, here we are most interested in players' changing knowledge as a game proceeds. The language allows us to make these issues more precise. For instance, if a player is certain now that after some move takes place φ is the case, then after that move, is she still certain that φ is the case? In other words, does the following formula hold under all circumstances?

$$K_i[a]p \to [a]K_ip.$$

The answer is negative for most of us. I know that I am boring after drinking – but it does not follow (unfortunately) that after drinking, I know that I am boring. The interchange axiom is only plausible for actions without 'epistemic side-effects'. And the converse implication can be refuted similarly. In general, dynamic-epistemic logic has no significant interaction axioms at all for knowledge

and action. If such axioms hold, this must be due to special features of the situation, such as special powers of agents qua observation or memory, or special features of the communicative relationship between agents.

5.2.3. Example. (games versus general dynamic-epistemic models). Imperfect information games themselves do satisfy a special axiom. The tree structure is common knowledge, and players cannot be uncertain about it. This is expressed by the following axiom – where M is the union of all available moves m in the game, and $m^{\cup}$ is the converse relation of m:

$$\langle K\rangle p \to \langle (M \cup M^{\cup})^*\rangle p \qquad (\#)$$

The effect of (#) can be stated as a modal frame correspondence. Epistemically accessible worlds are reachable from the root via sequences of moves:

5.2.4. Fact. (#) is true on a frame iff, for all s, t, if $s \sim_i t$, then $s(M \cup M^{\cup})^* t$.

Using this condition, every general model for a modal-epistemic language can be unraveled to a tree of finite action sequences in the usual modal fashion, with uncertainties $\sim_i$ between X, Y just in case $last(X) \sim_i last(Y)$. It is not hard to see that the map from sequences X to worlds $last(X)$ is then a bisimulation for the whole combined language.

Without this constraint, we get 'misty games' ([Höt03]), where players need not know what their moves are or what sort of opponent they are dealing with. This broader setting is quite realistic for planning problems. We return to it at the end of this chapter.

Axioms for perfect agents In the same correspondence style, the above knowledge-action interchange law really describes a special type of agent. To see this, we first observe that

5.2.5. Fact. $K_i[a]p \to [a]K_i p$ corresponds to the relational frame condition that for all s, t, u, if $sR_a t$ & $t \sim_i u$, then there is a v with $s \sim_i v$ & $vR_a u$.

This condition says that new uncertainties for an agent are always grounded in earlier ones. The equivalence can be proved, e.g. by appealing to the Sahlqvist form of this axiom. Incidentally, this and further observations about the import of axioms may be easier to understand using the equivalent existential versions, here: $\langle a\rangle\langle K\rangle p \to \langle K\rangle\langle a\rangle p$.

Precisely this relational condition was identified in [Ben01] as a natural version of players having *Perfect Recall* in the game-theoretic sense: They know their own moves and also remember their past uncertainties as they were at each stage. The actual analysis is slightly more complex in the case of games. First, consider

nodes where it is the player's turn: then $K_i[a]p$ implies $[a]K_ip$ for the same action a. Perfect Recall does not exclude, however, that moves by one player may be indistinguishable for others, and hence at another player's turn, $K_i[a]p$ implies merely that $[b]K_ip$ for some indistinguishable action b. But there are more versions of perfect recall in game theory. Some allow players uncertainty about the number of moves played by their opponents. [Bon04] has an account of such variants in essentially our correspondence style, now including a temporal operator into the language.

5.2.6. Remark. A similar analysis works for the converse dynamic-epistemic axiom $[a]K_ip \to K_i[a]p$, whose frame truth demands a converse frame condition of 'No Learning', stating essentially that current uncertainty relations remain under identical actions (cf. [FHMV95]). We will encounter this principle in a modified form in Section 5.3.

Agents with Perfect Recall also show special behaviour with respect to their knowledge about complex plans, including their own strategies.

5.2.7. Fact. Agents with Perfect Recall validate all dynamic-epistemic formulas of the form $K_i[\sigma]p \to [\sigma]K_ip$, where σ is a knowledge program.

Proof. By induction on programs. For knowledge tests $(K_i\varphi)?$, we have $K_i[(K_i\varphi)?]p \leftrightarrow K_i(K_i\varphi \to p)$ in dynamic logic, and then $K_i(K_i\varphi \to p) \leftrightarrow (K_i\varphi \to K_ip)$ in epistemic S5, and $(K_i\varphi \to K_ip) \leftrightarrow [(K_i\varphi)?]K_ip$ in dynamic logic. For the program operations of choice and composition, the inductive steps are obvious, and program iteration may be dealt with as repeated composition.□

This simple observation implies that an agent with Perfect Recall who knows what a plan will achieve will also know about these effects halfway through, when some part of his strategy has been played and only some remains. Again, this is not true for all types of agent. This is only one of many delicate issues that can be raised about players' knowledge of their strategies. Indeed, a knowledge statement about *objects*, like 'knowing one's strategy', has aspects that cannot be expressed in our formalism at all. We leave this for further elaboration elsewhere.

Axioms for imperfect agents But there are other types of agents! At the opposite of Perfect Recall, there are agents with bounded memory, who can only remember a fixed number of previous events. Such players with 'bounded rationality' are modelled in game theory by restricting them to strategies that can be implemented by some finite automaton (cf. [OR94]). [Ben01] considers the most drastic form of memory restriction, to just the last event observed. We will call them *memory-free* agents. This kind of agent will be our guiding example of epistemic limitations in this chapter.

In modal-epistemic terms, memory-free agents satisfy a Memory Axiom:

$\langle a\rangle p \to U[a]\langle K\rangle p$ *MF*

This involves extending our language with a *universal modality* $U\varphi$ stating that φ holds in all worlds. The technical meaning of *MF* is as follows.

5.2.8. Claim. *The axiom MF corresponds to the structural frame condition that, if* sR_at *and* uR_av*, then* $v \sim_i t$*.*

Thus, nodes where the same action has been performed are indistinguishable to memory-free agents. Reformulated in terms of knowledge, the axiom becomes $\langle a\rangle K_i p \to U[a]p$. This says that the agent can only know things after an action which are true wherever the action has been performed. Therefore, memory-free agents know very little indeed! We will study their behaviour further in Section 5.4. For now, we return to perfection.

5.3 Update for perfect agents

Imperfect information trees merely provide a static record of what uncertainties players are supposed to have at various stages of a game. And then we have to think of some plausible scenario which might have produced these uncertainties. One general mechanism of this kind is provided by *update logics* for actions with epistemic import. Recall Definition 2.5.5 from Chapter 2, where the product rule says that uncertainty among new states can only come from existing uncertainty via indistinguishable actions. That simple mechanism covers surprisingly many forms of epistemic update. [Ben03], [Dit05], [DHK07], [BGP07] and many other recent publications provide introductions to update logics and the many open questions one can ask about them.

The same perspective may now be applied to imperfect information games, where successive levels correspond to successive repetitions of the sequence

$\mathcal{M}$, $\mathcal{M} \times \mathcal{A}$, $(\mathcal{M} \times \mathcal{A}) \times \mathcal{A}$, ...

The result is an obvious tree-like model $\mathit{Tree}(\mathcal{M}, \mathcal{A})$, which may be infinite.

5.3.1. Example. (propagating uncertainty along a game). The following illustration is from [Ben01]. Suppose we are given a game tree with admissible moves (preconditions will be clear immediately). Let the moves come with epistemic uncertainties encoded in an action model, shown in Figure 5.2. Then the imperfect information game can be computed with levels as shown in Figure 5.3:

Now enrich the modal-epistemic language with a dynamic operator

$\mathcal{M}, s \models \langle \mathcal{A}, a\rangle\varphi$ iff $(\mathcal{M}, s) \times (\mathcal{A}, a) \models \varphi$.

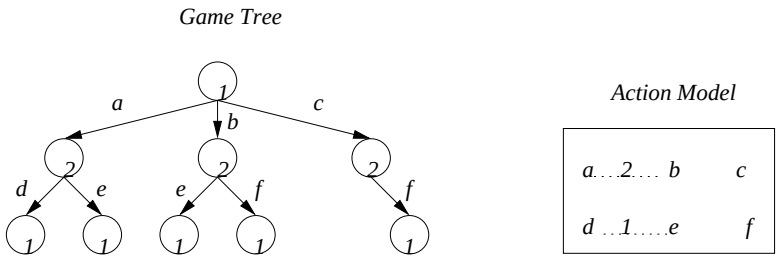

Figure 5.2: Game tree and action model

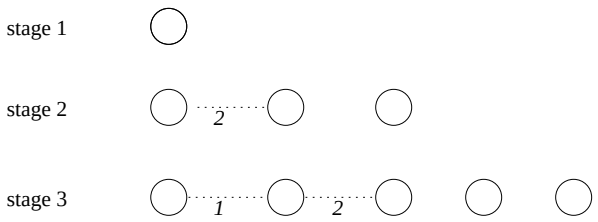

Figure 5.3: propagating uncertainty along a game

Then valid principles express how knowledge is related before and after an action. In particular, we have this key *reduction axiom*:

$$\langle \mathcal{A}, a\rangle\langle K\rangle\varphi \leftrightarrow (PRE(a) \wedge \bigvee\{\langle K\rangle\langle \mathcal{A}, b\rangle\varphi\colon\ a \sim_i b \textit{ for some } b \textit{ in } \mathcal{A}\}).$$

Such laws simplify reasoning about action and planning: We can reduce epistemic properties of later stages to epistemic information about the current stage. From left to right, this axiom is the earlier Perfect Recall, but now with a twist compared with earlier formulations. If an agent cannot distinguish certain actions from the actual one, then those may show up in his epistemic alternatives. The opposite direction from right to left is the No Learning principle. But it does not say that agents can never learn, only that no learning is possible for them among indistinguishable situations by using actions that they cannot distinguish.

The preceding logical observations show that product update is geared toward special agents, viz. those with Perfect Recall. The fact that the reduction axiom is valid shows that perfect memory must have been built into the very definition. And it is easy to see how. The two clauses in defining the new relation $(s,\ a) \sim_i (t,\ b)$ give equal weight to

(a) $s \sim_i t$: past states representing the 'memory component',

(b) $a \sim_i b$: options for the newly observed event.

Changes in this mechanism will produce other 'product agents' by assigning different weights to these two factors (see Section 5.5). But first, we determine the essence of product update from the general perspective of Section 5.2. The following result improves a theorem in [Ben01].

Abstract characterization of product update Consider a tree-like structure $\mathcal{E}$ with possible events (or actions) and uncertainty relations among its nodes, which can also verify atomic propositions p, q, ... The only contrast with a real tree is that we allow a bottom level with multiple roots. Nodes X, Y, ... are at the same time finite sequences of events, and the symbol $^\cap$ expresses concatenation of events. Intuitively, we think of such a tree structure $\mathcal{E}$ as the possible evolutions of some process – for instance, a game. A particular case is the above model $Tree(\mathcal{M}, \mathcal{A})$ starting from an initial epistemic model $\mathcal{M}$ and an action model $\mathcal{A}$, and repeating product updates forever. Now, the preceding discussion shows that the following two principles are valid in $Tree(\mathcal{M}, \mathcal{A})$, which can be stated as general properties of a tree $\mathcal{E}$. They represent Perfect Recall and 'Uniform No Learning', respectively:

PR If $X^\cap(a) \sim_i Y$, then $\exists b \exists Z : Y = Z^\cap(b)$ & $X \sim_i Z$.

UNL If $X^\cap(a) \sim_i Y^\cap(b)$, then $\forall U, V : if\ U \sim_i V$, then $U^\cap(a) \sim_i V^\cap(b)$, provided that $U^\cap(a), V^\cap(b)$ both occur in the tree $\mathcal{E}$.

Moreover, the special nature of the preconditions in product update, as definable conditions inside the current epistemic model, validates one more abstract constraint on the tree $\mathcal{E}$:

BIS-INV The set $\{X \mid X^\cap(a) \in \mathcal{E}\}$ of nodes where action a can be performed is closed *under purely epistemic bisimulations* of nodes.

Now we have all we need to prove a converse representation result.

5.3.2. Theorem. *For any tree $\mathcal{E}$, the following are equivalent:*

(*a*) $\mathcal{E} \cong Tree(\mathcal{M}, \mathcal{A})$ *for some* $\mathcal{M}, \mathcal{A}$.

(*b*) $\mathcal{E}$ *satisfies PR, UNL, BIS-INV.*

Proof. From (a) to (b) is the above observation. Now, from (b) to (a). Define an epistemic model $\mathcal{M}$ as the set of initial points in $\mathcal{E}$ and copy the relations $\sim_i$ from $\mathcal{E}$. The action model $\mathcal{A}$ contains all possible actions occurring in the tree, where we set

$$a \sim_i b \quad \text{iff} \quad \exists X \exists Y : X^\cap(a) \sim_i Y^\cap(b).$$

We also need to know that the *preconditions* $PRE(a)$ for actions a are as required. For this, we use the well-known fact that in any epistemic model, any set of worlds that is closed under epistemic bisimulations must have a definition in the epistemic language – though admittedly, one allowing infinite conjunctions and disjunctions. The abstract setting of our result allows no further finitization of this definability.

Now, the obvious identity map F sends nodes X of $\mathcal{E}$ to corresponding states in the model *Tree* $(\mathcal{M}, \mathcal{A})$. First, we observe the following fact about $\mathcal{E}$ itself:

5.3.3. Lemma. *If $X \sim_i Y$, then length(X) = length(Y).*

Proof. If X, Y are initial points in $\mathcal{E}$, both their lengths are 0. Otherwise, suppose X has length $n{+}1$. By *PR*, X's initial segment of length n stands in the relation $\sim_i$ to a proper initial segment of Y whose length is that of Y minus 1. Repeating this observation peels off both sequences to initial points after the same number of steps. □

5.3.4. Claim. *$X \sim_i Y$ holds in $\mathcal{E}$ iff $F(X) \sim_i F(Y)$ holds in $Tree(\mathcal{M}, \mathcal{A})$.*

The proof is by induction on the common length of the two sequences X, Y. The case of initial points is clear by the definition of $\mathcal{M}$. As for the inductive steps, consider first the direction $\Rightarrow$. If $U^{\frown}(a) \sim_i V$, then by *PR*, $\exists\, b \exists\, Z : V = Z^{\frown}(b)$ & $U \sim_i Z$. By the inductive hypothesis, we have $F(U) \sim_i F(Z)$. We also have $a \sim_i b$ by the definition of $\mathcal{A}$. Moreover, given that the sequences $U^{\frown}(a)$, $Z^{\frown}(b)$ both belong to $\mathcal{E}$, their preconditions as listed in $\mathcal{A}$ are satisfied. Therefore, in $Tree(\mathcal{M}, \mathcal{A})$, by the definition of product update, $(F(U), a) \sim_i (F(Z), b)$, i.e. $F(U^{\frown}(a)) \sim_i F(Z^{\frown}(b))$.

As for the direction $\Leftarrow$, suppose that in $Tree(\mathcal{M}, \mathcal{A})$ we have $(F(U), a) \sim_i F(Z), b)$. Then by the definition of product update, $F(U) \sim_i F(Z)$ and $a \sim_i b$. By the inductive hypothesis, from $F(U) \sim_i F(Z)$ we get $U \sim_i Z$ in $\mathcal{E}(*)$. Also, by the given definition of $a \sim_i b$ in the action model $\mathcal{A}$, we have $\exists\, X \exists\, Y : X^{\frown}(a) \sim_i Y^{\frown}(b)(**)$. Taking $(*)$ and $(**)$ together, by *UNL* we get $U^{\frown}(a) \sim_i Z^{\frown}(b)$, provided that $U^{\frown}(a)$, $V^{\frown}(b) \in \mathcal{E}$. But this is so since the preconditions $PRE(a)$, $PRE(b)$ of the actions a, b were satisfied at $F(U), F(Z)$. This means these epistemic formulas must also have been true at U, V – so, given what $PRE(a)$, $PRE(b)$ defined, $U^{\frown}(a)$, $V^{\frown}(b)$ exist in the tree $\mathcal{E}$. □

This result is only one of a kind, and its assumptions may be overly restrictive. In many game scenarios, preconditions for actions are not purely epistemic, but rather depend on what happens over time. E.g. a game may have initial factual announcements – like the Father's saying that at least one child is dirty in the puzzle of the Muddy Children. These are not repeated, even though their preconditions still hold at later stages. Describing this requires preconditions $PRE(a)$ for actions a that refer to the temporal structure of the tree $\mathcal{E}$, and then the above invariance for purely epistemic bisimulations would fail. Another strong assumption is our use of a single action model $\mathcal{A}$ that gets repeated all the time in levels $\mathcal{M}$, $(\mathcal{M} \times \mathcal{A})$, $(\mathcal{M} \times \mathcal{A}) \times \mathcal{A}$, ... to produce the structure $Tree(\mathcal{M}, \mathcal{A})$. A more local perspective would allow different action models $\mathcal{A}_1$, $\mathcal{A}_2$, ... in stepping from one tree level to another. And an even more finely-grained view arises if single moves in a game themselves can be complex action models. In the rest of this paper, for convenience, we stick to the single-model view.

5.4 Update logic for bounded agents

Limitations on information processing The information-processing capacity of agents may be bounded in various ways. One of these is 'external': Agents may have restricted powers of observation. This kind of restriction is built into the definition of action models, with uncertainties for agents – and the product update mechanism of Section 5.3 reflects this. Another type of restriction is 'internal': Agents may have bounded memory. Agents with Perfect Recall had limited powers of observation but perfect memory. At the opposite extreme we find memory-free agents who can only observe the last event, without maintaining any record of what went on before. In this section, we explore this extreme case.

Characterizing types of agent In the preceding, agents with Perfect Recall have been described in various ways. Our general setting was the tree $\mathcal{E}$ of event sequences, where different types of agents i correspond to different types of uncertainty relation $\sim_i$. One approach was via *structural conditions* on such relations, such as *PR*, *UNL*, and *BIS-INV* in the above characterization theorem. Essentially, these three constraints say that

$$X \sim_i Y \quad \text{iff} \quad \mathit{length}(X) = \mathit{length}(Y) \text{ and } X(s) \sim_i Y(s) \text{ for all positions } s.$$

Next, these conditions also validated corresponding *axioms in the dynamic-epistemic language* that govern typical reasoning about the relevant type of agent. But thirdly, we can also think of agents as a sort of *processing mechanism*. Intuitively, an agent with Perfect Recall is a push-down store automaton maintaining a stack of all past events and continually adding new observations to the stack. Such a processing mechanism was provided by our representation theorem, viz. epistemic product update.

Bounded memory Another broad class of agents arises by assuming bounded memory up to some fixed finite number k of positions. In general trees $\mathcal{E}$, this makes two event sequences X, Y $\sim_i$-equivalent for such agents i iff their last k positions are $\sim_i$-equivalent. In this section we only consider the most extreme case of this, viz. *memory-free agents* i:

$$X \sim_i Y \quad \text{iff} \quad \mathit{last}(X) \sim_i \mathit{last}(Y) \text{ or } X = Y = \text{the empty sequence} \qquad (\$)$$

Agents of this sort only respond to the last-observed event. In particular, their uncertainty relations can now cross between different levels of a game tree: They need not know how many moves have been played. Perhaps contrary to appearances, such limited agents can be quite useful. Examples are *Tit-for-Tat* players in the iterated Prisoner's Dilemma which merely repeat their opponents' last move ([Axe84]), or *Copy-Cat* players in game semantics for linear logic which

can win 'parallel disjunctions' of games $G \vee G^d$ ([Abr96]). Incidentally, these are players with a hard-wired *strategy*: a point that we will discuss below. It is easy to characterize such agents in terms similar to what we did with Perfect Recall.

5.4.1. FACT. An equivalence relation $\sim_i$ on $\mathcal{E}$ is memory-free in the sense of (\$) if and only if the following two conditions are satisfied:

PR^-	If $X^\frown(a) \sim_i Y$, then $\exists\, b \sim_i a \exists\, Z : Y = Z^\frown(b)$.
UNL^+	If $X^\frown(a) \sim_i Y^\frown(b)$, then $\forall\, U, V : U^\frown(a) \sim_i V^\frown(b)$, provided that $U^\frown(a)$, $V^\frown(b)$ both occur in the tree $\mathcal{E}$.

Proof. If an agent i is memory-free, its relation $\sim_i$ evidently satisfies PR^- and UNL^+. Conversely, suppose that these conditions hold. If $X \sim_i Y$, then either X, Y are both the empty sequence, and we are done, or, say, $X = Z(a)$. Then by PR^-, $Y = U(b)$ for some $b \sim_i a$, and so $last(X) \sim_i last(Y)$. Conversely, the reflexivity of $\sim_i$ plus UNL^+ imply that, if the right-hand side of the equivalence (\$) holds, then $X \sim_i Y$. □

It is also easy to give a characteristic modal-epistemic axiom for this case. First, we set the following

$$a \sim_i b \quad \text{iff} \quad \exists X \exists Y : X^\frown(a) \sim_i Y^\frown(b).$$

5.4.2. FACT. The following equivalence is valid for memory-free agents:

$$\langle a \rangle \langle K \rangle \varphi \leftrightarrow (PRE(a) \& E \bigvee_{b \sim_i a} \langle b \rangle \varphi).$$

Here $E\varphi$ is an additional *existential modality* saying that φ holds in at least one node. This axiom looks at first glance like the Perfect Recall axiom of Section 3, but note that there is no epistemic modality $\langle K \rangle$ on the right-hand side of the equivalence. Also, this new axiom implies axiom *MF* from Section 5.2, assuming that basic actions are partial functions.

5.4.3. REMARK. (reduction axioms for an existential modality). Once the static description language gets extended, to restore the harmony of an update logic, one should also extend the dynamic update reduction axioms with a clause for the new operator. E.g., returning to Section 5.3, the following reduction axiom is valid for standard product update:

$$\langle \mathcal{A}, a \rangle E\varphi \leftrightarrow (PRE(a) \wedge E \bigvee \langle \mathcal{A},\, b \rangle \varphi \textit{ for some } b \textit{ in } \mathcal{A}).$$

The process mechanism: finite automata The processor of memory-free agents is a very simple *finite automaton* creating their correct $\sim_i$ links:

> States of the automaton: all equivalence classes $X^{\sim_i}$
> Transitions for actions a: $X^{\sim_i}$ goes to $(X^{\cap}(a))^{\sim_i}$

There are only finitely many states since we had only finitely many actions in the game tree $\mathcal{E}$. The transitions are well-defined, since by the No Learning assumption UNL^+, if $X \sim_i Y$, then $X^{\cap}(a) \sim_i Y^{\cap}(a)$. The automaton starts in the equivalence class of the empty event sequence. Repeating transitions, it is easy to see that

> When the automaton is given the successive members of an event sequence X as input, it ends in state $X^{\sim_i}$.

In particular, $X \sim_i Y$ iff the automaton ends in the same state on both of these event sequences. Moreover, the combination of the conditions UNL^+ and PR^- on memory-free agents tells us something about the special type of automaton that suffices:

> All transitions a end in the same state (as $X^{\cap}(a) \sim_i Y^{\cap}(a)$ for all X, Y), and by PR^-, no transition ends in the initial state.

Let us call such automata *rigid*. They only have states for the last-observed event, and such states will even coincide when the events are not epistemically distinguishable for the agent.

5.4.4. Fact. Memory-free agents are exactly those whose uncertainty relation is generated by a rigid finite-state automaton.

Of course, more complex finite automata can have more differentiated responses to observed events a, up to some fixed finite number of cases.

5.4.5. Remark. (automata theory). Connections with automata theory, in particular the Nerode representation of finite automata recognizing regular sets of event sequences, are found in [BC03]. The above framework can be extended with more general preconditions for game actions referring to time, by generalizing to the action/test automata used for propositional dynamic logic in [HKT00].

Strategies and automata The preceding automata for bounded agents are reaction devices to incoming observations. But it is also tempting to think of automata as generators of behaviour – in particular, as specific *strategies*. The latter view is more in line with the usual treatment of our motivating examples, like *Tit-for-Tat* or *Copy-Cat*. A strategy for player i in a game is a function assigning moves to turns for i, these moves are responses to *other players'* actions.

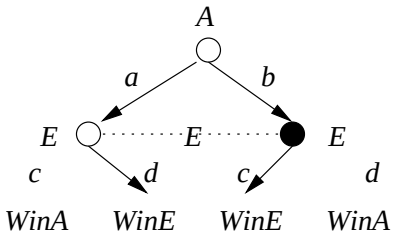

Figure 5.4: Winning strategy

This is easily visualized in game trees $\mathcal{E}$. E.g., player $\boldsymbol{E}$'s winning strategy in the game of Section 5.2 looks as shown in Figure 5.4.

But the reflection in finite automata will be a little different then, as players do not respond to a last action if played by themselves (these are 'non-events' for the purpose of a strategy). Thus, the usual automaton for *Tit-for-Tat* encodes actions by the agent itself as *states*, while actions by the opponent are the true observed events, shown in Figure 5.5.

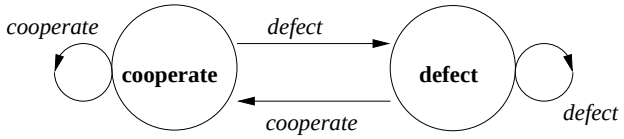

Figure 5.5: Tit for Tat

We do not undertake an integration of the two sorts of finite automata here. Either way, the simplicity of such automata for agents and their strategies may also be seen by considering the special syntactic form of memory-free strategies as simple knowledge programs in the dynamic-epistemic language.

This concludes our discussion of memory-free agents per se. To highlight them even more, we add a few contrasts with agents with Perfect Recall.

Differences in what agents know Memory-free agents i know less than agents with Perfect Recall. The reason is that their equivalence classes for $\sim_i$ tend to be larger. E.g., *Tit-for-Tat* only knows she is in two of the four possible matrix squares (*cooperate, cooperate*) or (*defect, defect*). But amongst many other failures, she does not know the accumulated score at the current stage. It is also tempting to say that memory-free agents can only run very simplistic strategies. But this is not quite right, since any knowledge program makes sense for all agents. The point is just that certain knowledge conditions will evaluate differently for both. E.g., a Perfect Recall agent may be able to act on conditions like "action a has occurred twice so far", which a memory-free agent can never execute, since she can never know that the condition holds. Thus the difference is rather in the number of non-equivalent available uniform strategies and the successful behaviour guaranteed by these.

5.4.6. Example. Consider the following game tree for an agent $\boldsymbol{A}$ with perfect information, and a memory-free agent $\boldsymbol{E}$ who only observes the last move.

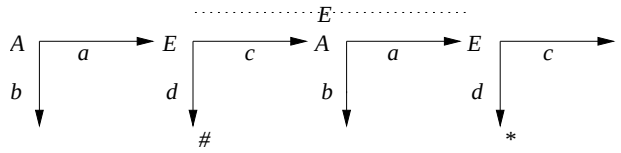

Figure 5.6: How memory-free agents may suffer

Suppose that outcome # is a bad thing, and $*$ a good thing for $\boldsymbol{E}$. Then the desirable strategy "play d only after you have seen two a's" is unavailable to $\boldsymbol{E}$ – while it is available to a player with Perfect Recall.

Another difference between Perfect Recall agents and memory-free agents has to do with what they know about their *strategies*. We saw that an agent with Perfect Recall for atomic actions also satisfies the key implication

$K_i[\sigma]p \to [\sigma]K_ip$, when σ is any complex knowledge program.

By contrast, the *MF* Memory Axiom

$$\langle a\rangle p \to U[a]\langle K\rangle p.$$

does not 'lift' to arbitrary knowledge programs instead of the single action a. To see this, it suffices to look at the case of a choice program $a \cup b$. Our eventual reduction version

$$\langle a\rangle\langle K\rangle\varphi \leftrightarrow (PRE(a) \& E \bigvee_{b \sim_i a} \langle b\rangle\varphi).$$

is a bit harder to generalize, because we would first have to analyze what it means to be indistinguishable from a complex action.

Memory and time A good way of making differences between agents more explicit is the introduction of a richer language. So far, we have mostly looked at a purely epistemic language for preconditions and an epistemic language with forward action modalities for describing updates or general moves through a game tree. With such a language, some of the intuitive distinctions that we want to make between different agents cannot be expressed. E.g., suppose that there is just one initial world s and one action, the identity Id, which always succeeds:

$$s \qquad (s, Id) \qquad ((s, Id), Id) \qquad ...$$

Thus, each horizontal level contains just one world. In this model, the uncertainty lines for Perfect Recall agents and memory-free agents are different. The latter see all worlds ending in Id as indistinguishable, whereas product update for the

former makes all worlds different. Nevertheless, agents know exactly the same purely epistemic statements in each world. The technical reason is that all states are epistemically bisimilar, and composing the uncertainty lines for a player with bisimulation links makes no difference to what she knows. But intuitively, the Perfect Recall player should know how many actions have occurred, since her uncertainties did not cross levels. Now, if we want to let agents know explicit statements about where they are in the game, we can add the backward-looking *converse action modalities* mentioned in Section 5.2. Then an agent knows, e.g., that two moves have been played if she knows that two consecutive converse actions are possible, but not three. Thus, a temporal dynamic-epistemic language is more true to what we would want to say intuitively about players and their differences. Moreover, this language can also express more complex preconditions for actions, resulting in the definability of a much broader range of strategies (cf. [Rod01], [BP06] and [BGP07]).

5.4.7. Remark. (backward-looking update). A backward-looking temporal language also enriches update logic. Our reduction axioms so far were forward-looking analysis of *preconditions*, reducing what agents know after an action has taken place to what they knew before. What about converse reduction axioms of the following form, say:

$$\langle a^{\cup}\rangle\langle i\rangle\varphi \leftrightarrow (PRE_{a^{\cup}} \& E \bigvee_{b^{\cup} \sim_i a^{\cup}} \langle b^{\cup}\rangle\varphi)?$$

These are related to *postconditions* for actions a: The strongest that we can say when a was performed in a world satisfying φ is that $\langle a^{\cup}\rangle\varphi$ must hold. Such postconditions are known to be impossible to define, even for simple public announcements, in the open-ended total universe of all epistemic models. But things are more controlled in our trees $\mathcal{E}$ which fix the previous history for any current world. In that case, we can convert at least earlier full commutativity axioms like the interchange of $\langle a\rangle\langle K\rangle$ and $\langle K\rangle\langle a\rangle$ to backward-looking versions. For more discussions, again we refer to [Yap06].

A final caveat This discussion has been somewhat impressionistic. In particular, it is easy to *over-interpret* our formal models in terms of 'knowledge talk'. At any given state, the bare fact is that an agent i has the set of all its $\sim_i$ alternatives. Depending on how *we* describe that set, we attribute various forms of knowledge to the agent. But most of these are just correlations – like when we say that *Tit-for-Tat* knows that it is in a 'cooperative' state. Such a description need not correspond to any *representational attitude* inside the agent. This mismatch is a limitation of epistemic logic in general, and over-interpretation occurs just as well for agents with Perfect Recall. These are triggered by possibly complex 'horizontal' knowledge conditions $K\varphi$ referring to the current tree level in structures like $\mathcal{E}$ or $Tree(\mathcal{M}, \mathcal{A})$. But we, as outside observers, may identify these as

equivalent to simple assertions about the past of the process, such as "action a has occurred twice". And even when we use the above richer temporal language, this still need not imply matching richer representations inside the agent.

5.5 Creating spectra of agents by modulating product update rules

Toward a spectrum of options Perfect Recall agents and memory-free agents are two extremes with room in the middle. Using the automata of Section 5.4, one might define update for progressively better informed k-bit agents having k memory cells, creating much great diversity. By contrast, agents with Perfect Recall seemed the natural children of product update. But even here there is room for alternative stipulations! The following type of agent is closely related to the memory-free ones discussed before.

Forgetful updaters As we saw in Section 5.3, product update for new uncertainties mixed a memory factor (viz. uncertainty between old states) and an observation factor (viz. uncertainty between actions). Agents might weigh these differently. A memory-free agent, by necessity, gives weight 0 to the past. If updating agents only remember their last action, how do they update their information? Here is a simple new definition. We drop the memory factor when defining product models $\mathcal{M} \times \mathcal{A}$, and set:

$$(x, a) \sim_i (y, b) \quad \text{iff} \quad a \sim_i b!$$

Thus, new uncertainty comes only from uncertainty about observed actions. Just as before, this leads to a valid *reduction axiom*:

5.5.1. Fact. The following equivalence is valid with forgetful update:

$$\langle \mathcal{A}, a\rangle\langle K\rangle\varphi \leftrightarrow (PRE(a) \wedge E \bigvee \langle \mathcal{A}, b\rangle\varphi\text{: } a \sim_i b \textit{ for some } b \textit{ in } \mathcal{A}).$$

As before, to restore the harmony of the complete system, we also need a reduction axiom for the new modality E, which turns out to be

$$\langle \mathcal{A}, a\rangle E\varphi \leftrightarrow (PRE(a) \wedge E \bigvee \langle A,\ b\rangle\varphi \textit{ for some } b \textit{ in } \mathcal{A}).$$

And it is also possible to give an abstract characterization of forgetful updaters by modifying the main theorem of Section 5.3.

In the original version of this chapter, it was suggested that forgetful updaters are precisely the memory-free agents of Section 5.4. But as was pointed out by Josh Snyder (personal communication), this seems wrong. Consider the following scenario. A forgetful updater is uncertain between world s with p and world t with $\neg p$. There are two possible actions:

- a with precondition: $p \wedge \neg Kp$,

- b with precondition: $Kp \vee (\neg p \wedge \neg K\neg p)$.

Let the actual actions be a, b in that order. Then the successive product updates for forgetful updaters are

(i) from $\{s,\ t\}$ to $\{(s,\ a),\ (t,\ b)\}$, without an uncertainty link, so the agent knows that p in the actual world $(s,\ a)$, whereas he knows that $\neg p$ in the unrelated world $(t,\ b)$

(ii) from $\{(s,\ a),\ (t,\ b)\}$ to $\{((s,\ a),\ b)\}$, since neither a nor b can be performed in $(t,\ b)$.

But in that final model, the agent still knows that p, even though a memory-free agent would not know p because she would be uncertain between $((s,\ a),\ b)$ and $(t,\ b)$. [Sny04] has a solution for this by modifying product update so as to keep all worlds around, whether or not preconditions of actions are satisfied, while redefining uncertainty relations in some appropriate fashion. Another option may be the addition of suitable 'copy actions' that keep earlier sequences alive at later levels. We will come back to these two proposals in Chapter 6.

The upshot of this discussion is that forgetful updaters are not the same as our earlier memory-free agents, although they are close. In the remainder of this section, we mention some other modulations on product update that create different types of agents.

Probabilistic modulations Letting agents give different weights to memory and observation in computing a new information state is an idea from a well-known tradition preceding modern update logics, viz. inductive logic and Bayesian statistics. Different agents or 'inductive methods' differ in the weight they put on experience versus observation. To implement this perspective in update logics, we need a *probabilistic* version of product update, as first defined in [Ben03], and later developed in [BGK06].

Belief revision and plausibility update But staying closer to our qualitative setting, we can also give another natural example of diversity with a numerical flavour. In the theory of *belief revision*, it has long been recognized that agents may obey different rules, more conservative or more radical, when incorporating new information. Such rules are different options for computing new states on the basis of incoming evidence. Such diversity will even arise for agents with epistemic Perfect Recall, as we will now show.

In general, information update is a different mechanism from belief revision, but the two viewpoints can be merged. [Auc03] adds a function κ to epistemic models $\mathcal{M}$ and action models $\mathcal{A}$ which assigns *plausibility values* to states and

actions. Here $\kappa_i(v) > \kappa_i(w)$ means that agent i believes that world w is more plausible than world v. This allows us to define degrees of belief in a proposition as truth in all worlds up to a certain plausibility:

$$\mathcal{M},\, s \models B_i^{\alpha}\varphi \text{ iff } \mathcal{M},\, t \models \varphi \text{ for all worlds } t \sim_i s \text{ with } \kappa(t) \leq \alpha.$$

Incidentally, we can also define $B_i^{\alpha}\varphi$ as $K_i(\kappa_i^{\alpha} \to \varphi)$, provided we add suitable propositional constants κ_i^{α} to the language (cf. [Liu04]).

Next, plausibilities of actions indicate what an agent believes about what most likely took place. Computing the plausibility of a new state $(w,\, a)$ in a product model $\mathcal{M} \times \mathcal{A}$ requires some intuitive rule. Aucher himself proposes an 'addition formula' for κ-values, subtracting a 'correction factor':

$$\kappa_j'(w, a) = Cut_M(\kappa_j(w) + \kappa_j^*(a) - \kappa_j^w(PRE(a))).$$

Here *Cut* is a technical 'rescaling' device, and the correction $\kappa_j^w(PRE(a))$ is the smallest κ-value in $\mathcal{M}$ among all worlds $v \sim_i w$ satisfying $PRE(a)$.

A continuum of revision rules In our current perspective, we see this stipulation not as the unique update rule for plausibility but as a choice for a particular type of agent. Aucher's formula makes an agent 'eager' in the following sense: The factor for the last-observed action weighs just as heavily as that for the previous state, even though the latter might encode a long history of earlier beliefs. But we can easily create further diversity by changing the above formula into one with parameters λ and μ:

$$\kappa_j'(w, a) = \frac{1}{\lambda + \mu}(\lambda\kappa_j(w) + \mu\kappa_j^*(a)).$$

By changing values of λ and μ, we can distinguish many different types of agents. Diversity increases even further when we let agents assign different plausibility values to preconditions of actions. For a detailed discussion, see [Liu04].

5.5.2. Remark. (belief revision by bounded agents). It is also possible to use ideas from Section 5.3, and consider belief revising agents with bounded memory. For a more extensive study of belief revision by agents with bounded resources, we refer to [Was00], [ALW04], and [AJL07].

Coming to terms with belief revision, in addition to information update, is natural – also from our motivating viewpoint of games. After all, players of a game surely do not just update on the basis of observed past moves. They also revise their expectations about future actions of opponents. Further examples of this will arise in our final sections.

5.6 Mixing different types of agents

So far, we have looked at agent types separately. But agents live in groups, whose members may have different types. Turing machines might communicate with finite automata, and humans occasionally meet Turing machines, like their computers, or finite automata, like very stupid people. What makes groups of agents most interesting is that they *interact*. In this setting, a host of new questions arises – of which we discuss just a few.

Uncertainty and exploitation Do different types of agents know each other's type? There is an issue of definition first. What does it *mean* to know the type of another agent? One could think of this, e.g., as knowing that the agent satisfies all axioms for its type, as formulated in Sections 5.2, 5.3 and 5.4. But then, in imperfect information games, or the more general trees $\mathcal{E}$ studied above, the types of all agents are *common knowledge*, because these axioms hold everywhere in the tree. Introducing ignorance of types requires more complex structures in the sense of [Höt03]. Suppose that agent $\boldsymbol{A}$ does not know if his opponent is a memory-free agent or not. Then we need disjoint unions of game trees with uncertainty links between them. Indeed, this extension already arises when we assume that some agent i does not know the precise uncertainties of its opponent between i's actions. Consider the following example:

5.6.1. Example. The following situation is a simple variant of Example 5.1, pictured in Figure 5.7.

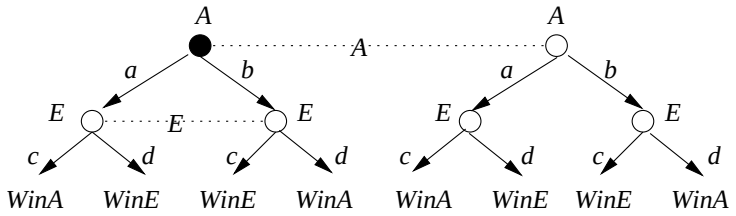

Figure 5.7: Ignorance of the opponent type

Right at the start of the game, agent $\boldsymbol{A}$ does not know whether $\boldsymbol{E}$ has limited powers of observation or not. In particular, note that the earlier axiom $\langle K\rangle p \to \langle (M \cup M^{\cup})^*\rangle p$ for imperfect information games fails here. The 'second root' toward the right is an epistemic alternative for $\boldsymbol{A}$, but it is not reachable by any sequence of moves.

Can an agent take advantage of knowing another agent's type? Of course. It would be tedious to give overly formal examples of this, since we all know this phenomenon in practice. Suppose that I know that after returning a serve of mine, you always step toward the middle of the court. Then passing you all the

time along the outer line is a simple winning strategy. A more dramatic scenario of this sort occurs in the movie "Memento" about a man who has lost his long-term memory and has fallen into the hands of unscrupulous cops and women. But *must* a memory-free agent do badly against a more sophisticated epistemic agent? That depends on the setting. E.g., memory-free *Tit-for-Tat* managed to win against much more sophisticated computer programs ([Axe84]). But even this does not do justice to the complexity of interaction!

Learning and revision over time In practice, we may not know the types of other agents and may need to *learn* them. Such learning mechanisms are themselves a further source of interesting epistemic diversity, as is pointed out in [Hen01] and [Hen03]. In general, there is no guarantee at all that a learning method will reveal the type of an opponent. Evidently, observing a finite number of moves can never tell us for certain whether we are playing against an agent with Perfect Recall or against a finite-state automaton with a large finite memory beyond the current number of rounds played so far. But there is a weaker sense of learning that may be more relevant here. We may enter a game with certain hypotheses about the agents that we are playing against. And such hypotheses can be updated by observations that we make as time goes by. E.g. I can *refute* the hypothesis that you are a memory-free agent by observing different responses to the same move of mine at different stages of the game. Or, I can have the justified hypothesis that you are memory-free, and one observed response to a move of mine then reveals a part of your fixed strategy.

Two kinds of update Intuitively, the game situations just described go beyond the information and plausibility update of Sections 5.3, 5.4, 5.5. But to arrive at a more definite verdict, one has to separate concerns. The above questions involve many general issues about update that arise even without diversity of agents. For instance, learning about one's opponent's type is akin to the well-known question of learning one's opponent's strategy. Types may be viewed as sets of strategies, so learning the type amounts to some useful intermediate reduction in the strategic form of the game. In what follows, we will illustrate a few issues in a concrete scenario.

5.6.2. Example. Consider the following game of perfect information. Suppose that $\boldsymbol{A}$ knows that $\boldsymbol{E}$ is memory-free: What does it take him then to find out which particular strategy $\boldsymbol{E}$ is running? See Figure 5.8.

This scenario illustrates the danger in discussing these matters. For, if $\boldsymbol{A}$ *knows* that $\boldsymbol{E}$ is memory-free, the latter fact is true, and hence, at her second turn, $\boldsymbol{E}$ can never play d, since she has already played c in response to b in order to get there at all. So, we can only sensibly talk about *beliefs* here. In the simplest case, these can be modelled as

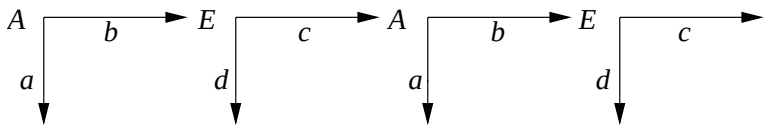

Figure 5.8: Finding out about types and strategies

subsets of all runs of the game from now on,

viz. those future runs which the agent takes to be most likely. Thus, $\boldsymbol{A}$'s belief would rule out the 'non-homogeneous run' for $\boldsymbol{E}$ in this game, even though further observation might refute the belief, forcing $\boldsymbol{A}$ to revise. Now, *belief revision* means that, as the game is played and moves are observed, this set of most plausible runs gets modified. E.g., suppose that $\boldsymbol{E}$ in fact plays d at her first turn. Then the hypothesis that she was memory-free seems vindicated, and we also know part of her strategy. But this is again too hasty. We have not tested any global assertion about her strategy, precisely because the game is over, and we have no means of observing what $\boldsymbol{E}$ would have done at her second turn.

Thus, we must be sensitive to distinctions like 'predicting what *will* happen' versus 'predicting what *would* happen' in some stronger counterfactual sense. Hypotheses about one's opponents' type are of the latter sort, and they may be harder to test. The representation of alternative scenarios and suitable update mechanisms over these need not be the same in both cases. In particular, we might need two kinds of update mechanisms. One is the local computation of players' uncertainties at nodes of the game concerning facts and other players' information, as described by the earlier product update and plausibility update. The other is the changing of the global longer-term expectations about strategies over time by observing the course of the game.

5.6.3. Remark. (local versus global update?). Despite the appealing distinction made just now, uncertainty about the future can sometimes be 'folded back' into local update. Consider any game of perfect information. Uncertainties about the strategy played by one's opponent may be represented in a new *imperfect information* game, whose initial state consists of all possible strategy profiles with appropriate uncertainty lines for players between these. Update on such a structure occurs as consecutive moves are played in the game, which can be seen as a form of public announcement ruling out certain profiles from the diagram. Likewise, belief revision becomes plausibility update on strategy profiles. For details, see [Ben04a] and [Ben04b].

Update can get even more subtle than this with learning global types. Consider the earlier Example 5.7 where $\boldsymbol{A}$ did not know if $\boldsymbol{E}$ had perfect information or not. How can $\boldsymbol{A}$ find out? If only moves are observed, we would have to say

that having just a single uncertainty line for $\boldsymbol{A}$ between the real root and the 'pseudo-root' makes no sense. For, after move a is played, $\boldsymbol{A}$ has learnt nothing that would now enlighten him, so there should be an uncertainty line at the mid-level as well. But in another sense, $\boldsymbol{A}$ *has* learnt something! He now knows that $\boldsymbol{E}$ is uncertain, so he is in the game on the left. To make sense of this second scenario, we have to assume that introspection into $\boldsymbol{A}$'s epistemic state also counts as an update signal.

We leave matters here. What we hope to have shown is that diversity of agents raises some interesting issues, while sharpening our intuitions about the required mix of update and revision in games. In particular, instead of theorizing about abstract revision mechanisms, a hierarchy of agent types suggests very concrete switching scenarios as our beliefs about a type get contradicted by events in the course of the game.

Merging update logic and temporal logic To make sense of the issues in this section, we need to introduce a richer framework than our dynamic-epistemic logic so far. We now need to maintain global hypotheses about behaviour of agents in future courses of the game, which can be updated as time proceeds. This temporal intuition reflects computational practice, as well as philosophical studies of agency and planning (cf. [BPX01]). It is also much like questions in standard game theory about predicting the future behaviour of one's opponents: 'rational', or less so. Technically, we think the best extension for this broader sort of update would be *branching temporal models* with a suitable language referring to behaviour over time (cf. [FHMV95], [PR03]). The above tree structures $\mathcal{E}$ can easily support such a richer language. [Ben04b] has a few speculations on update in such a temporal setting, and a recent exploration can be found in [BP06].

5.7 Conclusion

The point of this chapter is that diversity of agents is a fact of life, and moreover, that it is interesting from a logical point of view. Indeed, as we shall see in Chapter 6, one can even apply it to other logical core tasks, such as inference by more clever and more stupid agents. Technically, we have shown that it is easy to describe different kinds of epistemic agents in dynamic epistemic logics, and that this style of analysis matches well with information flow in extensive games of imperfect information.

Several interesting further questions arise now, and some of them have been taken up in the time since the paper [BL04] behind this chapter was first published. One line of extension is the further mathematical study of special patterns in arbitrary imperfect information games, viewed as trees of actions with epistemic uncertainties. In such a setting, our representation results may have more sophisticated versions for other kinds of behaviour. One could see this as pur-

suing the fine-structure of general models for dynamic-epistemic logic. Indeed, some generalized versions of our representation results are found in [BGP07]. [Ben07c] contains some further DEL-style game analysis in its section on extensive games. Further studies in this line are [Har04] on preference and player's powers in games, [Bru04] on epistemic foundations of game theory, [Ott05] on update in games concerning players' strategic intentions and preferences, [DZ07] and [Dég07b]. Also [Roy08] on the role of intentions and information dynamics in games, which develops connections with the philosophy of action.

Also, the results in this chapter suggest a richer temporal perspective, where belief changes do not just concern partial past observations, but also expectations about the future. This calls for a merge of temporal logic, dynamic-epistemic logic, and belief revision. For epistemic logic proper, this has been done in [BP06], with important protocol-based extensions in [BGP07]. Merging temporal logic with belief revision is done in [Bon07], and see [Zve07] for further elaboration. Branching temporal versions of dynamic belief revision in DEL-style have been explored in [Dég07a]. Related work includes [BHT06], [HT06], etc.

Similarly, the logical style of analysis presented here needs to be brought into contact with the ways in which game theorists study bounded rationality (cf. [OR94], [Rub98]). These tie in more with complexity-theoretic diversity in processing capacities of players, and/or computational difficulty of the games they are playing (cf. [Sev06]).

Finally, we think that interaction of diverse agents is a topic with many logical repercussions, of which we have merely scratched the surface. But this is a topic which will calls for a yet more general perspective on sources of diversity, to be presented in the next chapter.

Chapter 6

Diversity of Agents and their Interaction

6.1 Diversity inside logical systems

Logical systems seem to prescribe one norm for an "idealized agent". Any discrepancies with actual human behavior are then irrelevant, since the logic is meant to be normative, not descriptive. But logical systems would not be of much appeal if they did not have a plausible link with reality. And this is not just a matter of confronting one ideal norm with one kind of practical behavior. The striking fact is that human and virtual agents are not all the same: actual reasoning takes place in societies of diverse agents.

This diversity shows itself particularly clearly in *epistemic logic*. There have been long debates about the appropriateness of various basic axioms, and they have to do with agents' different powers. In particular, the ubiquitous modal Distribution Axiom has the following epistemic flavor:

6.1.1. Example. Logical omniscience: $K(\varphi \to \psi) \to (K\varphi \to K\psi)$.

Do rational agents always *know the consequences* of what they know? Most philosophers deny this. There have been many attempts at bringing the resulting diversity into the logic as a legitimate feature of agents. Some authors have used "awareness" as a sort of restriction on short-term memory ([FH85]), others have concentrated on the stepwise dynamics of making inferences ([Kon88], [Dun95]). A well-informed up-to-date philosophical summary is found in [Egr04].

The next case for diversity lies in a different power of agents:

6.1.2. Example. Introspection axioms: $K\varphi \to KK\varphi$, $\neg K\varphi \to K\neg K\varphi$.

Do agents *know when they know* (or *do not know*)? Many philosophers doubt this, too. This time, there is a well-established way of incorporating different powers into the logic, using different accessibility relations between possible worlds

in Kripke models. Accordingly, we get different modal logics: K, T, $S4$, or $S5$. Each of these modal logics can be thought of as describing one sort of agents. The interesting setting is then one of combinations. E.g., a combined language with two modalities K_1, K_2 describes a two-person society of introspectively different agents! This gives an interestingly different take on current logic combinations ([GS98], [KZ03]): the various ways of forming combined logics, by "fusions" $S5+S4$ or "products" $S5\times S4$, correspond to different assumptions about how the agents *interact* in an abstract sense. Effects may be surprising here. E.g., later on, in our discussion of memory-free agents, we see that knowledge of memory-free agents behaves much like "universal modalities". But in certain modal logic combinations, adding a universal modality drives up complexity, showing how the interplay of more clever and more stupid agents may itself be very complex...

Thus, we have seen how *diversity exists inside standard epistemic logic*, and hence likewise in doxastic logic. The purpose of this chapter is to bring to light some further sources of diversity in existing logics of information. Eventually, we would want to move from complaints about "limitations" and "bounds" to a positive understanding of how societies of diverse agents can perform difficult tasks ([GTtARG99]). In addition to identifying diversity of behavior, this also requires a study of *interactions* between different agents: e.g., how one agent learns the types of the agents she is encountering and makes use of such knowledge in communication. This chapter is structured as follows. Section 6.2 briefly identifies some further parameters of variation for agents beyond the well-known, and somewhat over-worked, concerns of standard epistemic logic. These are: powers of observation, powers of memory, and policies for belief revision. Section 6.3 then looks at dynamic epistemic logics of information update, showing how limited powers of observation for different agents are already accounted for, while we then add some new update systems which also describe varieties of bounded memory. Moving on to correcting beliefs on the basis of new information, Section 6.4 takes a parallel look at dynamic doxastic logics for belief revision, and shows how different revision policies can be dealt with inside one logical system. Section 6.5 is a brief summary of sources of diversity, and a transition to our next topic: that of interaction between different agents. In particular, Section 6.6 discusses several scenarios where different sorts of agent meet, involving identification of types of speaker (liars versus truth-tellers), communication with agents having different introspective powers, and encounters between belief revisers following different policies. We show how these can be dealt with in plausible extensions of dynamic-epistemic and dynamic-doxastic logics. Finally, in Section 6.7, we summarize, and pose some further more ambitious questions.

This chapter is based on existing literature, unpublished work in my Master's Thesis ([Liu04]) plus some new research in the meantime. We will mainly cite the relevant technical results without proof, and put them into a fresh story.

6.2 Sources of diversity

The diversity of logical agents seems to stem from different sources. In what follows, we shall mainly speak about "limitations", even though this is a loaded term suggesting "failure". Of course, the more cheerful reality is that agents have various resources, and they use these positively to perform many difficult tasks, often highly successfully.

Our epistemic axioms point at several "parameters" of variation of agents, and indeed, we already identified two of them:

(a) *inferential/computational power*: making all possible proof steps,

(b) *introspection*: being able to view yourself in "meta-mode".

One further potential parameter relevant to epistemic logic is the "awareness" studied by some authors ([FH85]), which suggests some resource like limited attention span, or short-term memory.

Next, consider modern dynamic logics of information, whose motivation sounds closer to actual cognitive practice. These also turn out to incorporate idealizations that suggest further parametrization for diversity. We start with the case of information update.

Consider the basics of *public announcement logic* (PAL): the event $!\varphi$ in this language means "the fact φ is truthfully announced". PAL considers the epistemic effects these announcement actions bring about. In addition to static epistemic axioms that invite diversity, here is a new relevant issue which merges only in such a dynamic setting. The following principle is crucial to the way PAL analyzes epistemic effects of public assertions, say, in the course of a conversation, or a sequence of experiments with public outcomes:

$$[!\varphi]K_a\psi \leftrightarrow \varphi \rightarrow K_a[!\varphi]\psi \quad \textit{Knowledge Prediction Axiom}$$

But the validity of this axiom presupposes several things, notably *Perfect Observation* and *Perfect Recall* by agents. The event of announcement must be clearly identifiable by all, and moreover, the update induced by the announcement only works well on a unique current information state recording all information received so far. This informal description is made precise in the detailed soundness proof for Knowledge Prediction Axiom in Section 6.3. Also, we will discuss this in the more general framework of "product update" for dynamic epistemic languages ([BMS98]). Thus, we have found two more parameters of diversity in logic. Agents can also differ in their powers of:

(c) *observation*: variety of agents' powers for observing current events,

(d) *memory*: agents may have different memory capacities, e.g., storing only the last k events observed, for some fixed k.

Can one deal with these additional forms of diversity inside the logic? As we will see, dynamic epistemic logic with product update can itself be viewed as a calculus of observational powers. And as to memory, [BL04] has shown how to incorporate this into dynamic epistemic logic (DEL) for memory-free agents, and we will extend their style of analysis below to arbitrary finite memory bounds.

The above four aspects are not the only places where diversity resides. Yet another source lies in *belief revision theory* ([AGM85]). Rational agents also revise their beliefs when incoming information contradicts what they believed so far. This scenario is different from the preceding one, as has been pointed out from the start in this area ([GR95]). Even for agents without limitations of the earlier sorts, there is now another legitimate source of diversity, viz. their 'learning habits' that create diversity:

(e) *revision policies*: varying from conservative to radical revision.

Different agents may react differently towards new information: some behave conservatively and try to keep their original beliefs as much as possible, others may be radical, easily accepting new information without much deliberation. However, these policies are not explicitly part of belief revision theory, except for some later manifestations ([Was00]). We will show in this chapter, following [Liu04], [BL07], how they can be brought explicitly into dynamic logic as well.

This concludes the list of parameters of diversity that we see in current dynamic-epistemic and dynamic-doxastic logics. It is important to mention that acknowledging this diversity inside logical systems is not a concession to the ugliness of reality. It is rather an attempt to get to grips with the most striking aspect of human cognition: despite our differences and limitations, societies of agents like us manage to cooperate in highly successful ways! Logic should not ignore this, but rather model it and help explain it. This chapter is a modest attempt at systematization toward this goal.

6.3 Dynamic logics of information update

Preliminaries in dynamic epistemic logic

To model knowledge change due to incoming information, a powerful current mechanism is dynamic epistemic logic, which has been developed intensively by [Pla89], [Ben96], [BMS98], [Ger99], [DHK07], etc. Since our discussions in this chapter will be based on DEL, we briefly recall its basic ideas and techniques.

6.3.1. Definition. An *epistemic model* is a tuple $\mathcal{M} = (S, \{\sim_a | a \in G\}, V)$ [1] such that S is a non-empty set of states, G is a group of agents, each $\sim_a$ is a

[1] We will sloppily write $\mathcal{M} = (S, \sim_a, V)$ when G is clear from the context.

binary epistemic equivalence relation, V is a map assigning to each propositional variable p in Φ a subset $V(p)$ of S.

We also have explicit models for our special citizens, the 'events'. Abstractly speaking, it has a similar structure as the epistemic model. Recall Definition 2.5.4 from Chapter 2.

The dynamic epistemic language is an extension of the one for standard epistemic logic. It is defined as follows

6.3.2. Definition. Let a finite set of propositional variables Φ, a finite set of agents G, and a finite set of events E be given. The *dynamic epistemic language* is defined by

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid [\mathcal{E}, e]\varphi$$

where $p \in \Phi$, $a \in G$, and $e \in E$.

As usual, $K_a\varphi$ stands for 'agent a knows that φ'. There are also new well-formed formulas of the type $[\mathcal{E}, e]\varphi$, which intuitively mean 'after event e takes place, φ will hold'. Here the $[\mathcal{E}, e]$ act as dynamic modalities. Thus, the expressiveness of the language is expanded in comparison with that of epistemic logic. One could also add the usual program operations of composition, choice, and iteration from propositional dynamic logic to the event vocabulary to deal with more complex situations like two events happening in sequence, choice of two possible events, and events taking place repeatedly. However in the current context, we will only consider a language without these operations.

6.3.3. Definition. Given an epistemic model $\mathcal{M} = (S, \{\sim_a \mid a \in G\}, V)$, we define $\mathcal{M}, s \models \varphi$ (formula φ *is true in* $\mathcal{M}$ at s) by induction on φ:

1. $\mathcal{M}, s \models \top$ always
2. $\mathcal{M}, s \models p$ iff $s \in V(p)$
3. $\mathcal{M}, s \models \neg\varphi$ iff not $\mathcal{M}, s \models \varphi$
4. $\mathcal{M}, s \models \varphi \wedge \psi$ iff $\mathcal{M}, s \models \varphi$ and $\mathcal{M}, s \models \psi$
5. $\mathcal{M}, s \models K_a\varphi$ iff for all $t : s \sim_a t$ implies $\mathcal{M}, t \models \varphi$.

In order to define the truth condition for the new formulas of the form $[\mathcal{E}, e]\varphi$, we need to define the product update model, again recall Definition 2.5.5 from Chapter 2. We can then add one more item for the truth definition of the formulas $[\mathcal{E}, e]\varphi$ to the above Definition 6.3.3:

6. $\mathcal{M}, s \models [\mathcal{E}, e]\varphi$ iff $\mathcal{M}, s \models PRE(e)$ implies $\mathcal{M}\times\mathcal{E}, (s, e) \models \varphi$.

Next, so called *reduction axioms* in DEL play an important role in encoding the epistemic changes. In particular, the following principle describes knowledge change of agents following some observed event in terms of what they knew before that event takes place:

$$[\mathcal{E}, e]K_a\varphi \leftrightarrow PRE(e) \rightarrow \bigwedge_{f \in \mathcal{E}} \{K_a[\mathcal{E}, f]\varphi : e \sim_a f\}.$$

Intuitively, after an event e takes place the agent a knows φ, is equivalent to saying that if the event e can take place, a knows beforehand that after e (or any other event f which a can not distinguish from e) happens φ will hold. Such a principle is of importance in that it allows us to relate our knowledge after an action takes place to our knowledge beforehand, which plays a crucial role in communication and general interaction.

This concludes our brief review of dynamic epistemic logic. We are ready to move to more complex situations where different agents live and interact. Public announcement logic is the simplest logic which is relevant here, as it describes agents who communicate via public assertions. This is the special case of DEL in the sense that the event model contains just one single event. The precondition of $!\varphi$ boils down to the fact that φ is true, as we will see in the formulas in the next section. In this chapter, for easy understanding, we use simple variants of PAL to motivate our claims, though we also consider a few scenarios using full-fledged DEL with a general mechanism of product update.

Public announcement, observation, and memory

First, we recall the complete axiom system for public announcement.

6.3.4. Theorem. ([*Pla89*][*Ger99*]). *PAL is axiomatized completely by the usual laws of epistemic logic plus the following reduction axioms:*

$(!p)$. $[!\varphi]p \leftrightarrow \varphi \rightarrow p$ *for atomic facts* p

$(!\neg)$. $[!\varphi]\neg\psi \leftrightarrow \varphi \rightarrow \neg[!\varphi]\psi$

$(!\wedge)$. $[!\varphi](\psi \wedge \chi) \leftrightarrow [!\varphi]\psi \wedge [!\varphi]\chi$

$(!K)$. $[!\varphi]K_a\psi \leftrightarrow \varphi \rightarrow K_a[!\varphi]\psi$.

Next, to introduce variety in *observation*, we need to assume a set of possible announcements $!\varphi$, $!\psi$, ... where an agent a need not be able to distinguish all of them. This uncertainty can be modelled by a simple event model with equivalence relation $\sim_a$ between statements which a cannot distinguish. The following example illustrates the difference in agents' powers of observation:

6.3.5. Example. Two agents a and b are traveling in Amsterdam and they want to visit the Van Gogh Museum. But they do not know whether Tram Line 5 goes there. A policeman said 'Tram Line 5 goes to the Van Gogh Museum'. a heard it, but b did not, as she was attracted by a Street musician who was playing her favorite song. So a learned something new, but b did not. Taking p to denote 'Tram Line 5 goes to the Van Gogh Museum', the state model and event model are depicted as follows:

Figure 6.1: State model and event model

The dotted lines express the epistemic uncertainties. The black nodes stand for the actual world and the actual event. The update leads to the following model:

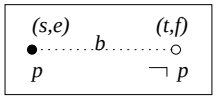

Figure 6.2: b is still uncertain

Note that at this stage a can distinguish between the two possible worlds, but b is still uncertain. There is diversity in observation!

The following principle – a special case of the above general DEL reduction axiom – then describes what agents know on the basis of partial observation:

6.3.6. Fact. The following reduction axiom is valid for agents with limited observation power:

$$[!\varphi]K_a\chi \leftrightarrow (\varphi \rightarrow \bigwedge_{!\psi \sim_a !\varphi} K_a[!\psi]\chi)$$

But there is another natural source of diversity, not dealt with by either PAL or DEL. As we have seen in the previous section, *Perfect Recall* assumes that agents can remember all the events that have happened so far. But in reality agents usually have bounded memory, and they can only remember a fixed number of previous events. It is much harder in PAL to model memory difference because the world elimination update procedure shifts agents to ever more informed states. To show the difficulty, consider the following example concerning *memory-free* agents which only acknowledge distinctions made by the last announcement, having no record of things further back in their past:

6.3.7. Example. Memory-free agent a is uncertain about p at first. Then p is announced, and afterwards, an "idle" action Id takes place. Then a should not know p any more since she does not remember anything. But here is what our standard update would do:

According to Definition 2.5.5, the model changes in the following way:

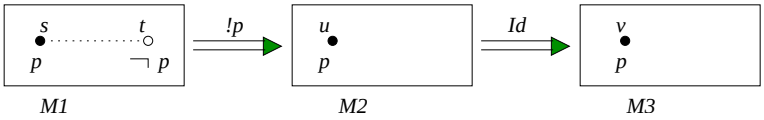

Figure 6.3: Memory-free agent remembers!

There are two possible worlds in the original model $\mathcal{M}_1$, the agent a is uncertain about p. After p is announced, we get $\mathcal{M}_2$. Since p does not hold at the world t, the action $!p$ only executes successfully at the world s, so we have only one world u in the model. Intuitively, after the announcement of p, agent a should now know that p, and indeed this holds in $\mathcal{M}_2$. Next, the Id action happens, which executes successfully everywhere. We get $\mathcal{M}_3$, abbreviating (u, Id) as v. Intuitively, once the action Id has been performed, the memory-free agent a should no longer know whether p, because she already forgot what had happened one step ago, and she should be uncertain again whether p. But in our model sequence, the agent a *knows* p. This is counter-intuitive!

Here is the reason. Standard product update eliminates possible worlds. Therefore, it is impossible to retrieve uncertainty links between worlds that have disappeared. There are several ways of amending this, and two proposals will be presented in detail later in this section. For the moment, we sketch one simple option suggested by [BL04]. First, we need to reformulate PAL update as in [BL07] to never eliminate worlds. The idea is to let announcements $!\varphi$ cut all links between φ-worlds and $\neg\varphi$-worlds, but otherwise, keep all worlds in. In this semantic perspective, the resulting "unreachabilities" between worlds represent the information that agents have so far. One way of describing a memory-restricted agent is then as having forgotten part or all of these "link removals". In the most extreme case, a memory-free agent will only consider distinctions caused by the last announcement – while reinstating all indistinguishability links that had been cut before. (Thus, longer sequences of announcements make no sense for such an agent: it is the last thing said which counts.) In particular, in this update scenario, worlds may also become indistinguishable again: a direct modelling of 'forgetting'. Forgetful agents like this do not satisfy the earlier reduction axiom $(!K)$, as is shown in the following example.

6.3.8. Example. Consider the two model changes depicted in Figure 6.4.

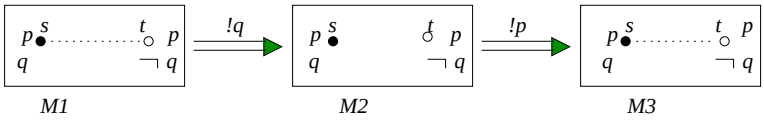

Figure 6.4: Reduction axiom fails

There are two possible worlds, s and t in $\mathcal{M}_1$, p and q hold at s, p and $\neg q$ hold at t. After q is announced, we get a new model $\mathcal{M}_2$, in which there is no uncertainty link between s and t. Then we have $(\mathcal{M}_2, s) \models p \to K_a(p \to q)$, i.e. $(\mathcal{M}_2, s) \models p \to K_a[!p]q$. After that, p is announced, and we have $\mathcal{M}_3 \not\models K_a q$, since the agent forgot $!q$ already. We look back at $\mathcal{M}_2$: $(\mathcal{M}_2, s) \not\models [!p]K_a q$. The reduction axiom does not hold!

With these examples in mind, what is the dynamic epistemic logic of forgetful agents? We will merely discuss a few issues. [BL04] gives the following modified reduction axiom, which trades in a knowledge operator after a dynamic modality for a universal modality $U\varphi$: 'φ is true in all worlds, accessible or not':

$$[\mathcal{E}, e]K_a\varphi \leftrightarrow PRE(e) \to \bigwedge_{e \sim_a f \in \mathcal{E}} U[\mathcal{E}, f]\varphi.$$

This is based on their version of product update which models agents who forget everything except the last event observed by changing the product update rule to this stipulation:

$$(s, e) \sim'_a (t, f) \quad \text{iff} \quad e \sim_a f.$$

Incidentally, to make this work technically, the system also needs a reduction axiom for the universal modality, and it reads as follows:

$$[\mathcal{E}, e]U\varphi \leftrightarrow PRE(e) \to \bigwedge_{e \sim_a f \in \mathcal{E}} U[\mathcal{E}, f]\varphi.$$

Transposed to just the current setting of public announcements (i.e., event models with one publicly observable event), this yields the following principle for forgetful agents:

$$[!\varphi]K_a\psi \leftrightarrow \varphi \to U[!\varphi]\psi.$$

These principles show that it is quite possible to write dynamic-epistemic axioms for agents with bounded memory, in the same style as before. Next, as in [BL07], take the link-cutting variant of public announcements of φ. This amounts to using event models with two events $!\varphi$ and $!\neg\varphi$ which are distinguishable for all agents. Again, the reduction law for forgetful agents follows in a simple manner.

Nevertheless, modeling memory in dynamic epistemic logics raises additional issues, of which we merely mention one. Notice that the preceding K/U equivalence completely obliterates the accessibility structure of the epistemic model

that was modified by the last announcement. E.g., the forgetful agent will know that fact q holds after a public announcement of p iff (assuming that p holds) every p-world (whether accessible or not) was a q-world. This may be considered a drawback of the above approach. There appears to be an intuitive difference between (a) forgetting what events took place and (b) what initial situation one started from. An intuitive alternative, suggested by the preceding examples, might let the agent remember the initial model. Here is an illustration of the difference between the two perspectives.

6.3.9. Example. Let the starting model be the following (Figure 6.5), where one world has already become inaccessible (but it might still be accessible via epistemic links for other agents):

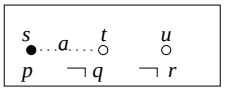

Figure 6.5: Initial model $\mathcal{M}_1$

Announcing $\neg r$ by public link cutting will leave this intact, we get the same picture with actual world s. Announcing $\neg q$ by public link cutting in the radical manner then would give us the following, as shown in Figure 6.6.

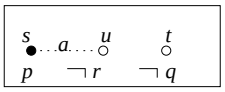

Figure 6.6: Announcing $\neg q$

But if the agent is supposed to remember the initial model, the outcome should be one where she knows that p is the case, see Figure 6.7.

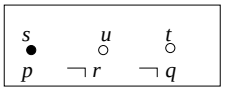

Figure 6.7: If the agent remembers...

Interestingly, implementing the latter less radical view of defective memory means that we have to keep track of *the initial model* $\mathcal{M}_1$, through long sequences of announcements. The reason is that there need not be enough information in $\mathcal{M}_1$'s successive modifications through updates to retrieve its original structure uniquely. Thus, while the behavior of agents with perfect memory may be described by just keeping track of the current epistemic model with all updates performed, the behavior of forgetful agents may require keeping track of a longer

history. This may sound paradoxical, but the point is that the latter book-keeping is to be done by the *modeler*, rather than the agent.

We will not formulate reduction axioms for our alternative version of bounded memory here. (Cf. the digression on epistemic temporal logic later in this section for some hints). Even at this somewhat inconclusive stage, however, we see that endowing agents with bounded memory can be achieved in principle.

Our overall conclusion is this: "Logic of public announcement" is actually a family of dynamic epistemic systems, with different update rules depending on the memory type of the agents, and correspondingly, different reduction axioms and reasoning styles.

Adding memory to product update

The previous section shows that the reduction axiom for knowledge under product update fails for memory-free agents. In this section we are going to propose a correct update rule for agents who have a bounded memory for the last observed events. By a *k-memory* agent, we mean an agent that remembers only the last k events before the most recent one. A 0-memory or memory-free agent does not recall anything; a 1-memory agent knows only what she learned from the last two actions, and so on. Modeling this diversity requires some care, witness the Example 6.3.7. As we mentioned, the difficulty there is that eliminating worlds is a form of hard-wired memory: worlds that have been removed do not come back, so one is 'forced to know'. To get this right in a more sensitive manner, we now present two proposals for product update with general memory-free agents. The first source for this is as follows:

6.3.10. Definition. ([Sny04]) Let an epistemic model $\mathcal{M} = (S, \sim_a, V)$ and an event model $\mathcal{E} = (E, \sim_a, PRE)$ be given. The *product update for memory-free agents* is $\mathcal{M} \times \mathcal{E} = (S \otimes E, \sim'_a, V')$ with:

(i) $S \otimes E = \{(s, e) : (s, e) \in S \times E\}$.

(ii) $(s, e) \sim'_a (t, f)$ iff ($\mathcal{M}$, s $\models PRE(e)$ iff $\mathcal{M}, t \models PRE(f)$) and $e \sim_a f$.

(iii) $V'(p) = \{(s, e) \in S \otimes E\colon\ s \in V(p)\}$.

Compared with the standard product update, item (i) in the above definition leaves out the precondition restriction. This keeps all worlds around. Item (ii) then defines the uncertainty relation on all worlds ('active', or not) in the new models. (iii) remains the same, and we will ignore this valuation clause henceforth. To understand this new definition, we look at the example again, now updating models according to the new definition, see Figure 6.8.

This is like Example 6.3.7 – but now, the original state model remains the same. According to Definition 6.3.10, we obtain a different model $\mathcal{M}_2$, abbreviating $(s, !p)$ as u and writing t as v. There is no uncertainty link between them. So

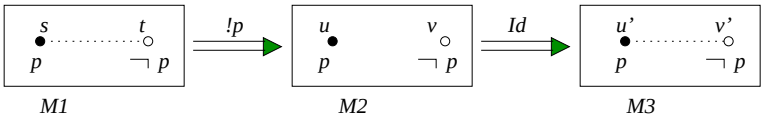

Figure 6.8: How memory-free agents update

the agent a knows that p in $\mathcal{M}_2$. Now the 'idle' identity event Id happens, and we get a new state model $\mathcal{M}_3$, abbreviating (u, Id) as u' and (v, Id) as v'. The agent a is uncertain whether p. This is what we expect for a 0-memory agent. [Sny04] also extended this proposal to the k-memory case.

Here, however, we also put propose an alternative for modelling forgetting, which seems closer to the workings of an actual memory store for agents. We introduce an auxiliary *copy action* $!C$ which always takes an old possible world into the new model with its reflexivity relation. Essentially it puts those worlds which were previously deleted into a stack, and makes sure agents can always retrieve them when needed.

6.3.11. Definition. ([Liu04]) Let an epistemic model $\mathcal{M} = (S, \sim_a, V)$ and an event model $\mathcal{E} = (E, \sim_a, PRE)$ be given. The *product update for memory-free agents* is $\mathcal{M} \times \mathcal{E} = (S \otimes E, \sim'_a, V')$ with:

(i) $S \otimes E = \{(s, e) \in S \times E\colon \mathcal{M}, s \models PRE(e)\}$.

(ii) For $e, f \neq !C$, $(s, e) \sim'_a (t, f)$ iff $e \sim_a f$.

To see how this new proposal works, we go back to the above example, but now update with an additional copy action:

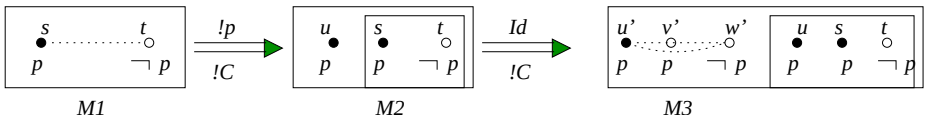

Figure 6.9: Update with copy actions

From the original model, by Definition 6.3.11, we get model $\mathcal{M}_2$, with a new state $(s, !p)$ abbreviated as u and two copied state s and t. The agent a then knows that p. To distinguish the new state and copied state, we put those copied ones in a rectangular box. After the Id action, similarly, we obtain the new model $\mathcal{M}_3$ with new states (u, Id) abbreviated as u', (s, Id) abbreviated as v', and (t, Id) abbreviated as w'. Again, $\mathcal{M}_3$ contains states that are copied from the previous model u, s and t. Again, the agent a is uncertain whether p. This idea is similar to the usual design of operation systems ([SGG03]), where the working memory does the jobs while carrying a stack of old information to be visited when necessary. [Liu04] has a more restrictive variant of the above definition copying

worlds only when necessary. This makes the above models less over-loaded.

Extending this approach, we get the following generalized update rule:

6.3.12. DEFINITION. ([Liu04]) Let $\mathcal{M}$ be an epistemic model, $\mathcal{E}_{-k}$ be the k-th event model before the most recent one $\mathcal{E}$. The *product update for k-memory agents* is $\mathcal{M} \times \mathcal{E}_{-k} \times \cdots \times \mathcal{E}_{-1} \times \mathcal{E} = (S \otimes E_{-k} \otimes \cdots \otimes E_{-1} \otimes E, \sim'_a, V')$ with:

(i) $S \otimes E_{-k} \otimes \cdots \otimes E_{-1} \otimes E = \{(s, e_{-k}, \ldots, e_{-1}, e) \in S \times E_{-k} \times \cdots \times E_{-1}$: $\mathcal{M} \otimes \mathcal{E}_{-k} \otimes \cdots \otimes \mathcal{E}_{-1}, (s, e_{-k}, \ldots, e_{-1}) \models PRE(e)\}$.

(ii) For $e_{-k}, \ldots, e_{-1}, e, f_{-k}, \ldots, f_{-1}, f \neq C!$,
$(s, e_{-k}, \ldots, e_{-1}, e) \sim'_a (t, f_{-k}, \ldots, f_{-1}, f)$ iff $e_{-k} \sim_a f_{-k}, \ldots, e_{-1} \sim_a f_{-1}$ and $e \sim_a f$.

Given this update rule, it is straightforward to find a complete dynamic logic in the earlier DEL format, but now for k-memory agents. Here we only consider the case in which $k = 1$, the uncertainty relation in the updated model is the above definition becomes:

For $e_{-1}, e, f_{-1}, f \neq C!, (s, e_{-1}, e) \sim'_a (t, f_{-1}, f)$ iff $e_{-1} \sim_a f_{-1} \& e \sim_a f$.

This is to say that a 1-memory agent cannot distinguish between two states in the new updated model, if and only if she cannot distinguish the two events that just took place, and neither the two events that had happened before. The reduction axiom for 1-memory agent is given in the following:

$$[\mathcal{E}, e_{-1}, e]K_a\varphi \leftrightarrow (PRE(e_{-1}) \wedge PRE(e) \rightarrow \bigwedge_{f_{-1}, f \in \mathcal{E}} \{K_a[\mathcal{E}, f_{-1}, f]\varphi : e_{-1} \sim_a f_{-1} \& e \sim_a f\}),$$

where e_{-1}, e, f_{-1}, f are not copy actions. Note that we have put two events that are relevant to 1-memory agents into the formula. Since copy actions function independently, we get a reduction axiom that is similar to the one we have for agents with perfect recall.

Of course, this is only the beginning of an array of further questions. In particular, we would like to have a more structured account of memory, as in computer science where we update data or knowledge bases. Update mechanisms are more refined there, referring to memory structure with actions such as information replacement ([Liu04]), where the agent would have a priority order in her database, so that she would know which old information should go to make room for the new. This is one instance of a more "constructive" syntactic approach to update, complementary to our abstract one in terms of model manipulation. Whether

our current semantic method or a syntactic one works better for finding agents' parameters of diversity is a question worth investigating.

Digression: Temporal Logic of Forgetful Agents
An alternative, and in some ways more concrete semantic framework for agents with memory bounds are branching tree models for epistemic-temporal logic ([BL04], [BP06]). Nodes in these models are finite sequences of events starting from the root of the tree, and epistemic indistinguishability relations between nodes model what agents have and have not been able to observe. In this setting, the epistemic accessibility relation for a forgetful agent recording just the last event simply becomes this:

$$X \sim_a Y \text{ iff } last(X) = last(Y), \text{ where } last(Z) \text{ is the last event in } Z.$$

As pictured in the following,

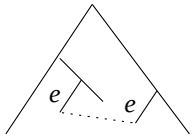

Figure 6.10: Epistemic relations in event trees

Now, the earlier dynamic epistemic reduction axioms become epistemic temporal principles, with indexed modalities $[e], \langle e \rangle$ have their obvious meaning referring to extensions $X^\frown e$ of the current node X. E.g., forgetful agents satisfy the following equivalence:

$$[e]K_a\varphi \leftrightarrow \langle e \rangle \top \rightarrow U[e]\varphi.$$

Note again how this trades an epistemic knowledge modality for a universal modality, as in the earlier examples in the previous subsection. The reason is that any node $X^\frown e$ in the temporal tree is epistemically related to any other node $Y^\frown e$. It is now straightforward to find similar principles for the knowledge of agents whose memory retains the last k observed events, as described above.

The total effect of these reduction axioms is as follows. Knowledge modalities are traded in for modal-temporal ones, as the accessibility relation is temporally definable in the model, and hence the epistemic-temporal language reduces to a purely temporal one. [BP06] use this reduction to show that the logic of memory bounded agents is computationally simpler than that of agents with perfect recall. This ends our digression.

This section has identified two new parameters for dynamic updating agents: powers of *observation* and powers of *memory*. DEL as it stands already provides a way of modelling the former, while we have shown how it can also be modified

to accommodate agents with bounded memory. We consider these two additional phenomena at least as important from an epistemological viewpoint as the usual themes of inferential power and introspection, generated by the earlier static phase of logical theorizing. Of course, as we have noted already, there is no need now to assume that all agents have the same powers. Indeed, our systems can describe the interplay of bounded and idealized agents, including ways in which one might exploit the other.

In the following section, we move to an extension of DEL treating one further crucial aspect of agents' cognitive behavior, when 'things get rough'.

6.4 Diversity in dynamic logics of belief change

Information flow and action based upon it is not always a matter of just smooth update. Another striking phenomenon is the way agents correct themselves when encountering evidence which contradicts their beliefs so far. *Belief revision theory* describes what happens when an agent is confronted with new information which conflicts her earlier beliefs. It has long been acknowledged that there is not one single logical rule for doing this. Indeed, different policies toward revising beliefs, from more 'radical' to more 'conservative' all fall within the compass of the famous AGM postulates.

In this chapter, however, we take another approach inspired by dynamic epistemic logic. First, on the static side, we follow the common idea that beliefs are modelled by so-called *plausibility relations* between worlds, making some epistemically accessible worlds more plausible than others. Agents believe what is true in the most plausible worlds – and the same thinking may also be used to define their conditional beliefs. In this setting, one can then view belief revision on the analogy of the preceding update paradigm, viz. as a mechanism of *change in plausibility relations*. To see this, here is a concrete example of how this can be implemented technically.

Belief revision as changing plausibility relations

One common policy for belief revision works as follows:

6.4.1. Example. ([Ben07a]) ($\Uparrow$) Radical revision
$\Uparrow P$ is an instruction for replacing the current ordering relation $\leq$ between worlds by the following: all P-worlds become better than all $\neg P$-worlds, and within those two zones, the old ordering remains.

Note that the $\neg P$-worlds are not eliminated here: they move downward in plausibility. This reflects the fact that we may change our mind once more on the basis of further information. $\Uparrow P$ is one famous policy for belief revision, corresponding to an 'eager response', or a 'radical revolution', or 'high trust' in the

source of the information. But there are many other policies in the literature. Another famous one would just place the best P-worlds on top, leaving the further order unchanged. A more general description of such different policies can be given as definable ways of changing a current plausibility relation ([BL07], [Rot06]). Once we have such a definition for a policy of plausibility change, the corresponding dynamic logic for belief revision can be axiomatized completely in DEL style. Here is the result for the policy of radical revision:

6.4.2. Theorem. ([*Ben07a*]) *The dynamic logic for radical revision* ($\Uparrow$) *is axiomatized completely by an axiom system KD45 on the static models, plus the following reduction axioms*

$(\Uparrow p)$. $[\Uparrow \varphi]p \leftrightarrow p$

$(\Uparrow \neg)$. $[\Uparrow \varphi]\neg\psi \leftrightarrow \neg[\Uparrow \varphi]\psi$

$(\Uparrow \wedge)$. $[\Uparrow \varphi](\psi \wedge \chi) \leftrightarrow [\Uparrow \varphi]\psi \wedge [\Uparrow \varphi]\chi$

$(\Uparrow B)$. $[\Uparrow\varphi]B\psi \leftrightarrow (E\varphi \wedge B([\Uparrow\varphi]\psi|\varphi)) \vee (\neg E\varphi \wedge B[\Uparrow\varphi]\psi)$

In the last axiom, E is the existential modality, dual to the earlier universal modality U. The symbol | denotes a conditional belief, and it means: 'given that'. Van Benthem's full system also has complete reduction axioms for conditional beliefs, thereby solving the notorious 'Iteration Problem' of AGM theory. This reduction axiom for the new beliefs shows precisely the doxastic effects of the chosen policy.

In the same style, one can also axiomatize other belief revision policies. For instance,'conservative revision' may be defined as follows: $\uparrow\varphi$ replaces the current ordering relation by the following: *the best φ-worlds come on top, but apart from that, the old ordering remains.* [Ben07a] presents a complete set of reduction axioms for this second policy as well. When put together, the result is a dynamic logic of belief revision which describes interactions between agents with different policies, using operator combinations such as, say, $[\Uparrow\varphi][\uparrow\psi]\chi$, which says that after a radical revision with φ followed by a conservative revision with ψ, the proposition χ holds.

All this is still qualitative. But the earlier product update mechanisms also admit of a more refined quantitative version, describing agents' attitudes in a more detailed numerical manner, and allowing for further polices of changing these fine-grained beliefs. In the next subsection, we will briefly show how.

Belief revision as changing plausibility values

Following [Spo88], a κ-ranking function was introduced in [Auc03] to extend DEL with numerical beliefs. A κ-ranking function maps a given set S of possible

worlds into the class of numbers up to some maximum Max. The numbers can be thought of as denoting degree of surprise. 0 denotes 'unsurprising', 1 denotes 'somewhat surprising', etc. κ represents a plausibility grading of the possible worlds, in other words, degree of beliefs.

6.4.3. Definition. A *doxastic epistemic model* is a tuple $\mathcal{M} = (S, \sim_a, \kappa_a, V)$, where S, $\sim_a$ and V are defined as usual, and the plausibility function κ_a ranging from 0 to some upper limit Max is defined on all worlds.

6.4.4. Definition. A *doxastic epistemic event model* is a tuple $\mathcal{E} = (E, \sim_a, \kappa_a^*, PRE)$, with E, $\sim_a$ and PRE defined as usual, κ_a^* ranges from 0 to Max, defined on all events.

The κ_a^*-value describes the agent's detailed view on which event is taking place. With plausibilities assigned to states and events, 'graded beliefs' will change via a suitable rule for product update. Here is the quantitative key proposal in [Auc03], the first of its kind in the DEL-style literature:

$$\kappa'_a(s,e) = Cut_{Max}(\kappa_a(s) + \kappa_a^*(e) - \kappa_a^s(\varphi)),$$

where $\varphi = PRE(e)$, $\kappa_a^s(\varphi) = min\{\kappa_a(t) : t \in V(\varphi) \text{ and } t \sim_a s\}$, and

$$Cut_{Max}(x) = \begin{cases} x & \text{if } 0 \leq x \leq Max \\ Max & \text{if } x > Max. \end{cases}$$

While this system looks formidable, a simple more perspicuous version exists. It uses an epistemic-doxastic language with propositional constants to describe the plausibility change ([Liu04]):

6.4.5. Definition. The *epistemic-doxastic language* is defined as

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid q_a^\delta$$

where $p \in \Phi$, a set of propositions, $a \in G$, a set of agents, and δ is a κ-value in $\mathbb{N}$, q_a^δ are a special type of propositional constants.

The interpretation is as usual, but now with the following simple truth condition for the additional propositional constants:

$$(\mathcal{M}, s) \models q_a^\delta \quad \text{iff} \quad \kappa_a(s) \leq \delta.$$

The numerical update mechanism can now be defined quite simply by merely specifying the new κ-value in the product model $\mathcal{M} \times \mathcal{E}$. To keep our discussion simple, we use just the following stipulation:

6.4.6. Definition. (bare addition rule). The new plausibilities for pair-worlds (s, e) in product models are defined by the following rule:

$\kappa'_a(s,e) = \kappa_a(s) + \kappa^*_a(e)$.

In this setting, reduction axioms assume a particularly simple form:

6.4.7. THEOREM. ([*Liu04*]) *The complete dynamic logic of plausibility belief revision consists of the key reduction axioms in Theorem 6.3.4 plus the new:*

$$[!\varphi]q^{\delta}_a \leftrightarrow q^{\delta-\kappa_a(!\varphi)}_a.$$

More generally, different update functions will account for different numerical revision policies. If such an update rule is simply expressible, we can get a complete dynamic logic for it in the style of the preceding result, though mere subtraction may not work anymore.

Additional power of description is provided by yet another device, viz. numerical parameters weighing the contributions of various factors. To illustrate this additional diversity of behavior for agents, we now present an update rule which incorporates further 'degrees of freedom':

6.4.8. DEFINITION. ([Liu04]) Let agent a assign weight λ to world s, and weight μ to the event e. The *plausibility of the new world* (s,e) is calculated by the parametrized rule

$$\kappa'_a(s,e) = \frac{1}{\lambda+\mu}(\lambda\kappa_a(s) + \mu\kappa^*_a(e)) \quad (\natural).$$

Intuitively, κ gives a degree of belief. The two parameters λ and μ express the importance of the state information, and that of the action information, respectively. Their variations then describe a range of various agents. For instance, when μ=0, we get *highly conservative agents*, and the ($\natural$) rule turns into $\kappa'_a(s,e) = \kappa_a(s)$. This means that the agent does not consider the effect of the last-observed event at all. Of course, some normalization is needed here to make sure that the new value is still in $\mathbb{N}$ (cf. [Liu06b]). Similarly, when λ=0, the agents are *highly radical*, and $\kappa'_a(s,e) = \kappa^*_a(e)$. When $\lambda = \mu$, we get *'Middle of the Road agents'* who let plausibility of states and actions play an equally important role in determining the plausibility of the new state. We obtain *conservative agents* when $\lambda > \mu$ and *radical agents* when $\mu > \lambda$. In this manner, we have distinguished five types of agents in dynamic logic. For an even more general view of agents' behavior towards incoming information, see [Liu06b]. Summing up, we may regard our numerical update rule as a refinement of the qualitative dynamic logics for belief change in the previous subsection (cf. [Ben07a] and [BL07]).

6.4.9. REMARK. Another relevant comparison is with the probabilistic update semantics proposed in [BGK06]. There the system computes probability values for pairs (s,e) using weighted products of prior world probabilities, occurrence probabilities for the type of event occurring, and observation probabilities describing agents' access to it. We defer a more detailed comparison of our views on agents' processing diversity with qualitative and probabilistic update logics to another occasion.

Some further observations

Our treatment of belief revision provides a simple format of plausibility change, where different policies show naturally in the update rules for either plausibility relations or value constants, and their matching reduction axioms in the dynamic doxastic logic. Moreover, our treatment also goes beyond the standard *AGM* paradigm, in that more complex event models allow agents to doubt the current information in various ways. Here are a few further issues that come up in this setting, some conceptual, some technical.

First, doubting the current information might also make sense for *PAL* and *DEL* scenarios even without belief revision involved. It is easy to achieve this by simply adding further events to an event model, providing, say, a public announcement $!\varphi$ with a counterpart $!\neg\varphi$ with some plausibility value reflecting the strength of the "dissenting voice". Likewise, policies with weights for various factors in update make much sense in recently proposed dynamic logics of probabilistic update (cf. [Auc05], [BGK06]).

Incidentally, this *DEL* approach via modified event models for different policies may also suggest that we can *relocate* policies from "modified update rules" to "modified event models" with a standard update rule. This has to do with an important more general issue: are we describing single events of update or revision 'locally' without further assumptions about the long-term behavior of the agents involved, or are we witnessing different more 'global' types of agent at work? In the former case, the diversity is in the response, rather than the type. We must leave this issue, and a comparison between the pros and cons of the two stances to another occasion.

Finally, connecting Sections 6.3 and 6.4, revision policies and memory restrictions may not be that disjoint after all. Technically speaking, the update behavior of highly radical agents is similar to that of memory-free agents, as they simply take the new information without considering what happened before (of course, for different reasons). In other words, the event that takes place completely characterizes the "next" epistemic state of the agent. This seems to be related also to notions such as "only knowing" or "minimal knowledge" in [Lev90] and [HJT90]. This final observation also provides a further challenge: viz. unifying some of our parameters of diversity discussed so far.

6.5 From diversity to interaction

We have investigated many different sources of diversity, some visible in static logics, some in dynamic ones. Besides the old parameters from epistemic logic, namely computation and introspection ability, we have added several new aspects, i.e. observation power, memory capacity and revision policy. Our discussion has been mostly in the framework of dynamic epistemic logic and we have shown

how it is possible to allow for a characterization of diversity within the logic. To summarize, look at the following diagram consisting of the main components of dynamic epistemic logic:

Static language	*Epistemic model* $\mathcal{M}$
Dynamic language	*Event model* $\mathcal{E}$
Product update	*Model change* $\mathcal{M} \times \mathcal{E}$

In the preceding sections we have shown that the diversity of agents can be explicitly modeled in terms of these logical components. The following table is an outline of the sources we have considered:[2]

Component	Residence	Diversity
$\mathcal{M}$	relations between worlds	introspection
$\mathcal{E}$	relations between actions	observation
$\mathcal{M} \times \mathcal{E}$	update mechanism	memory, revision policy

As we can see from the table, by introducing parameters of variation in each component, we are able to describe diversity of agents inside the logic.

But recognizing and celebrating diversity is only a first step! The next important phenomenon is that diverse agents *interact*, often highly successfully. Describing this interaction raises a whole new set of issues. In particular, our logical systems can describe the behavior of various agents, but they cannot yet state in one single formula "that an agent is of a certain type" or describe what would happen when we encounter those different agents. And as they stand, they are even less equipped to describe the interplay of different agents in a compact illuminating way. Imagine, if you know the type of the agent that you are encountering right now, can you take advantage of that knowledge? Or how could you *learn* about the type of the agent? In the following section, we will explore a few of these issues, and show in how far our current logical framework can handle these phenomena – and what features need to be added.

6.6 Interaction between different agents

Interaction between different agents is a vast area of diverse phenomena, and so, we will only discuss a few scenarios. These will show how the earlier dynamic logics can deal with some crucial aspects - though they also quickly need significant extensions. Our examples cover: reliability of sources (truth-tellers versus liars),

[2]Note that we have not discussed the earlier-mentioned parameter of inferential/computational power for agents. A more syntax-oriented approach to this topic can be find in [AJL06] and [Jag06]. It seems possible to merge the models proposed there with ours, and [Ben08] contains some first proposals for combined inferential and observational updates.

meetings between more or less introspective agents, and interaction between belief revisers following different policies.

'Living with Liars': dynamic logics of agent types

In this section we are challenging one of the PAL assumptions, namely, that all the announcements are truthful. What would happen if the announcer is a liar? More generally, can we figure out whether the announcer is a liar or truth-teller? In the following we will focus on such issues and explore how we update our knowledge when encountering people who should be identified first. These questions also bring us to a well-known puzzle about liars and truth-tellers. Here we consider one of its variations, high-lighting the fact that knowing what type of agent you encounter makes life a lot easier:

6.6.1. Example. On a fictional island, inhabitants either always tell the truth, or always lie. A visitor to the island meets two inhabitants, Aurora and Boniface. What the visitor can do is ask questions to discover what he needs to know. His aim is to find out the inhabitants' type from their statements. The visitor asks a what type she is, but does not hear a's answer. b then says "a said that she is a liar". Can you tell who is a liar and who is a truth-teller?

One can try to figure out the answer to the puzzle by intuitive reasoning, but we will give a precise analysis in logical terms in what follows. To describe the situation with the relevant events, the salient fact is the agent-oriented nature of the communication. To bring this out, we first need to extend the language with notation for agent types:

6.6.2. Definition. Take a finite set of propositional variables Φ, and a finite set of agents G. Predicates $L(x)$, $T(x)$ and action terms $!\varphi_a$ are now added. The *dynamic epistemic agent type language* is defined by the rule

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid [\pi]\varphi \mid L(x) \mid T(x)$$
$$\pi := !\varphi_a$$

where $p \in \Phi$, and $a \in G$.

Here $L(a)$ is intended to express 'agent a is a Liar', and $T(a)$ expresses 'agent a is a Truth-teller'. In fact, for the above example, we only need one of these expressions, since the agent is either a liar or a truth-teller. So we can use $\neg L(a)$ to denote 'agent a is a Truth-teller'. Besides, we also want to express *who* executes some action. Accordingly, $!\varphi_a$ reads intuitively as 'an announcement of φ performed by agent a'. Next we enrich the structure of our models, to a first approximation, in the following structures with hard-wired known agent types:

6.6.3. Definition. We define new epistemic models as $\mathcal{M} = (S, \{\sim_a \mid a \in G\}, V, L, T)$, where L, T are two types of agents, Liars and Truth-tellers. Moreover, given some suitable event model $\mathcal{E}$, the *truth conditions* for the new well-formed formulas are the following:

1. $\mathcal{M}, s \models T(x)$ iff $x \in T$.
2. $\mathcal{M}, s \models L(x)$ iff $x \in L$.
3. $\mathcal{M}, s \models [!\varphi_a]\psi$ iff ψ holds at the world $(s, !\varphi_a)$ in the product model $\mathcal{M} \otimes \mathcal{E}$.

Clause 1 and 2 are simple, as we only have two types of agents here. In general, there may be a larger set of types $\{L_1, L_2, \ldots L_k\}$, and we would then need to introduce a type function τ such that $\tau(L_i) \subseteq G$, setting $\mathcal{M}, s \models L_i(x)$ iff $x \in \tau(L_i)$. Item 3, however, is incomplete as it stands! This is because we have not given a precise update rule for the new agent-oriented announcements, which would require suitable *preconditions* $\langle !\varphi_a \rangle \top$ for the event of agent a's saying that φ.[3] In order to state useful and precise preconditions, we will definitely need more information about agent types.

Consider the example again. Clearly, the reason why the visitor should first find out who belongs to what type of agent is that it immediately determines the way she judges the incoming information. Here is a general illustration:

Case One: The visitor b does not know whether p is true, but *she knows that the speaker a is a truth-teller*. In fact, p is the case, and a says 'p is the case', after which b updates her knowledge accordingly:

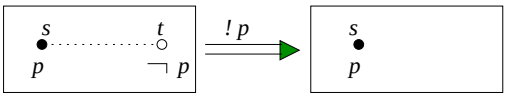

Figure 6.11: Telling the truth

Case Two: Next, the visitor b first does not know if p is true, but *she knows that a is a liar*. Now a says that 'p is not the case'. Agent b updates her knowledge with p instead of $\neg p$, see Figure 6.12:

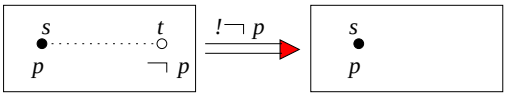

Figure 6.12: Lying

[3]Preconditions for agents' saying certain things may be related to their reliability according to the observing agent. Such a reliability judgment typically need not be publicly known. Thus, diversity of agents leads us to relax another idealization in standard *DEL* as defined earlier, viz. that preconditions of events are common knowledge.

These examples presuppose a definition of agent types, and how they affect preconditions for assertions. In the present scenario, these can be expressed more precisely in the following way:

(1) **truth-teller** $T(a) \to (\langle !\varphi_a \rangle \top \leftrightarrow \varphi)$

(2) **liar** $L(a) \to (\langle !\varphi_a \rangle \top \leftrightarrow \neg\varphi)$

Clause (1) says that a truth teller a can say exactly those things φ that are true. For the liar, this reverses.[4]

Even this simple stipulation has some interesting effects. E.g., no one can say that she is a liar, since our simple logic can formalize a version of the Liar Paradox, as stated in the following fact:

6.6.4. FACT. $\langle !L(a)_a \rangle \top$ does not hold in any case.

Proof. Suppose $\langle !L(a)_a \rangle \top$. There are two cases. Either a is a liar, $L(a)$, or a is a truth-teller, $T(a)$. In the first case, according to (2), we have $\langle !L(a)_a \rangle \top \to \neg L(a)$. Thus, in this case, we get $\neg L(a)$. But if a is a truth-teller, according to (1), we get $\langle !L(a)_a \rangle \top \to L(a)$ – and hence we have $L(a)$. This is a contradiction, and therefore, $\langle !L(a)_a \rangle \top$ does not hold. □

Incidentally, another take on our scenario might make it out to be about just single "lies and truths", rather than long-term liars and truth-tellers. This will not change our analysis here, but it would shift the emphasis in modeling from diversity of agents to what might be called *diversity of signals*. The latter tack is attractive, too, and sometimes simpler - but our main emphasis here is highlighting agent diversity in its own right.

Now as for interaction, we need to describe in general what agents would learn from communication if they knew the type of the other agent. To compute this, we can combine the information about agent types with the general rules of dynamic epistemic logic. For instance, even just minimal modal logic applied to the earlier type definitions yields the following principles:

(3) $K_b T(a) \to K_b(\langle !\varphi_a \rangle \top \leftrightarrow \varphi)$.

(4) $K_b L(a) \to K_b(\langle !\varphi_a \rangle \top \leftrightarrow \neg\varphi)$.

Using also the earlier reduction axioms for knowledge after events have taken place will generate further insights. Here are a few more valid principles about agents' changing knowledge in case a proposition is announced by a source whose type they know:

[4] See [BGP07] for a general account of more realistic conversational scenarios, where the current truth of a proposition need not imply that agents are automatically allowed to say it.

(6) $K_b T(a) \to ([!\varphi_a]K_b\varphi \leftrightarrow K_b[!\varphi_a]\varphi)$.

(7) $K_b L(a) \to ([!\varphi_a]K_b\neg\varphi \leftrightarrow K_b[!\varphi_a]\neg\varphi)$.

Of course, these principles are not yet a full-fledged account of messages. We have analyzed part of the information about the sender, but not yet the fact that it is a message from agent *a to agent b*. For logics of communication with such further aspects from the protocol perspective, we refer to [DW07].

Uncertainty about agent types

Still, the above is not all we need for our Island Puzzle. There, and also in real life, the types of agents encountered may be *unknown*! We need to represent that in our static and dynamic models. There are several ways of doing this. At the very least, the above predicates L, T will no longer be fixed once and for all for agents. They need to be made part of the specification of worlds, or events, so as to allow for uncertainty about them.

One proposal for modeling agent types (cf. [BGK06]) uses *pair events* of the form '(agent type, physical event)', say, "P is said by a truth-teller", or "P is said by a liar". Such abstract events are then epistemically indistinguishable if we can neither tell the agent types apart nor the actual observed events. However, in our analysis of the Island Puzzle, we do not need this rich format yet, since the conversation itself is about the types of agents, which makes things much easier. We therefore stick with a more ad-hoc format.

To model the original epistemic state of the visitor, see Figure 6.13 below. There is no information to indicate who is of what type, therefore, there are 4 possibilities in total, where for example the vertex (1, 1) represents the case in which a and b are both truth-tellers.

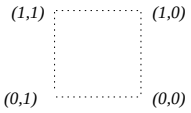

Figure 6.13: Initial model

Again, the dotted line denotes the visitor's uncertainty. Since the visitor does not hear what a says, there is no update for that.[5] Then b says "a said that she is a liar". Since we already noted the general truth that no one can say she is a liar, what b said about a is not true. So we conclude that b is a liar. This reasoning depends on the following principle, which follows from our agent type definition:

(5) $\varphi \wedge \langle!\neg\varphi_a\rangle\top \to L(a)$.

[5] In a more refined multi-agent scenario, there *would* be a product update for this event, as some higher-order knowledge about others changes – but we ignore this aspect here.

Meanwhile, we also know that a must have said that she is a truth-teller, since she was asked what type of agent she is, and there are only two possible answers. In this way, we (or the visitor to the island) split what b said into two statements: 'b is a liar' and 'a is a truth-teller'. To illustrate this more clearly, the update may be carried out in sequence, first with 'b is a liar', see Figure 6.14.

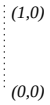

Figure 6.14: After knowing 'b is a liar'

And then with 'a is a truth-teller':

Eventually, we have obtained the required answer: Aurora is a truth-teller, while Boniface is a liar.[6]

Our analysis is in the same spirit as when one tries to figure out what kind of color a card has according to sequential announcements (cf. [Ben06a]). What is new here is that we no longer take any incoming information automatically as truthful. Instead, we first identify the type of agent who makes the statement, then we update our knowledge. Of course, this is only the beginning, since more complex scenarios would involve our updating our ideas about the *degree of reliability* of the source of our information.

The earlier valid principles about agents' changing knowledge when listening to speakers whose types they know easily extends to more complex event models with product events encoding uncertainty about agent types. The earlier general dynamic-epistemic reduction axioms will still work in this setting, when combined with preconditions for the different agent types.

Summing up, we have seen how an adequate account of different sources requires structured communicative events with agents explicitly indicated, explicit representations of agents' types, and a combination of general dynamic-epistemic reasoning principles with specific postulates about types of agent. In such a system, we can derive interesting principles about interaction between different agents. Of course, there are many more types of agent than just Liars and Truth-tellers, and Islands like the above are still logical paradise as compared to the real world. In particular, our views of the reliability of agents may change over time in subtle manners, calling for probabilistic information ([BGK06]). We will leave such further complications to future investigation.

[6]Strictly speaking, this is not quite right, since there is only one event of b's speaking, but we leave the formulation of one single update using our general product update mechanism to the reader.

A meeting between introspective and non-introspective agents

In this subsection we move to the perspective of the *addressee* instead of the *addressor* as investigated in the preceding scenario. Consider the following story.

6.6.5. Example. Two agents are sitting silently on a bench in the park. One of them, a, is non-introspective, but the other: b, is. The complete epistemic situation they find themselves in is depicted below – where the actual world is called s. The agents do not communicate with each other at first. Now, the aim for both of them is to find out which world is the real one as soon as possible. They have only one chance to receive new atomic information from some passer-by, and then they are ready to communicate with each other. What information should they get? What kind of communication should they engage in?

We picture the initial situation in the following diagram. As usual, all worlds are reflexive for each agent, but loops are omitted:

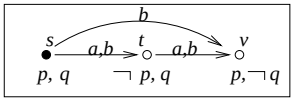

Figure 6.15: Initial situation

Here s is the real world where p and q are true. So, there are two possible atomic announcements one can make, either $!p$ or $!q$. Let's compare what will happen in these two cases. First, when q is truly announced by someone, the new model is pictured in Figure 6.16.

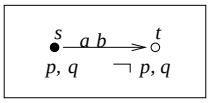

Figure 6.16: Two agents know the same

This new situation is symmetric between both agents. Both a and b are uncertain between s and t, and they do not know that s is the real world. And, given the symmetry, even if they communicate, it does not help, since they both know the same.

By contrast, once the fact p is announced, the new model becomes

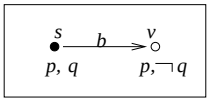

Figure 6.17: Two agents know differently

Here, the effect of this announcement is different for agent a and b! Agent b learns that p, but not that q. But agent a learns that both p and q, since she has no link to the world v. And she knows this is the real situation. Now a can inform b of q, so that b would also know q and realize that s is the real situation. What is going on here? How can the less-introspective agent a learn more? Do intellectuals need help from the man in the street to get their bearings?

We will just analyze what is going on here in terms of straight update. In terms of our epistemic models, a non-introspective agent may have fewer accessibility arrows than a corresponding introspective one, which means she is better informed, even though she does not reflect on this, and may not know everything she knows. Thus, additional information may help her more than her introspective companion.

To model the reasoning in such situations, as we have in the previous subsection, we can introduce agent types $I(a)$ and $NI(a)$ in the language to express that 'a is an introspective agent', and 'a is an anti-introspective agent', respectively, providing type definitions like the following:

(1) $I(a) \to (K_a\varphi \to K_aK_a\varphi)$.

(2) $NI(a) \to (K_a\varphi \to \neg K_aK_a\varphi)$.[7]

Clearly, because of their different introspective abilities, agents a and b may obtain quite different knowledge from what they learn. Intuitively, as we said already, the non-introspective agent even has an advantage in the above initial model, in that the following implication holds:

$$\mathcal{M}, s \models K_b\varphi \to K_a\varphi \quad (*).$$

But it is easy to think of settings where the knowledge of the agents would be incomparable. One can also analyze this type of situation more generically, using reduction axioms for informational events like before, leading to principles describing the interaction of the two agents such as the following:

(3) $NI(a) \wedge K_aI(b) \to ([!\varphi_c]K_b\psi \to K_a[!\varphi_c]\psi)$.

This is the static situation, looking at the agents separately. Of course, our scenario also illustrates another phenomenon, viz. how helpful agents which differ in their capacities may still inform each other, making the group consisting of both

[7]Note that one needs at least a non-normal logic to deal with anti-introspective agents. Since for instance the K-necessity rule $\vdash \varphi$, then $\vdash K_a\varphi$ itself presupposes certain positive introspection. It can lead to a contradiction. Moreover, given the definition (2), it is impossible to assume $K_aK_a\varphi \to K_a\varphi$, since we get $K_aK_a\varphi \to \neg K_a\varphi$ from (2). It would be interesting to investigate how far it is possible to model anti-introspective agents in modal logics.

agents together better informed than its members separately. Thus, our earlier observation that agents with different introspective powers lead to mere sums of modal logics $S4$ or $S5$ becomes just part of a more complex dynamic logic of what happens when they communicate.

Talking with different belief revisors

In our final scenario, we consider both information update and belief revision, and we also allow for diversity of both senders and receivers of information. Can our update models and their logics handle this? The following story is a bit contrived, but it highlights some realistic issues in everyday settings.

6.6.6. Example. Four agents live together, and their types are common knowledge. Agent a is a radical belief revisor, and b a conservative one. Agent c is a very trustworthy person, according to a and b, but d is less so. In the initial situation, there are three possible worlds s (the actual world), t, and v, as pictured in Figure 6.18 below, which also shows the valuation for the proposition letters. As for epistemic or doxastic relations, initially, a and b consider all three worlds possible, and they have the same plausibility ordering over them: v is most plausible, s is least plausible, t is in between. Moreover, c happens to know that p is the case, and d happens to know that q_2 is not the case. One can only speak after the other. Does this matter? Will both orders inform a, b equally well?

The original model may be depicted in Figure 6.18.

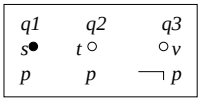

Figure 6.18: The original model: all agents believe the same.

Let us now suppose that d speaks first, truly, and says that p. Because of the different attitudes towards this new information, even though she acknowledges that d might be wrong, the radical (or more trusting) agent a will then change her plausibility ordering over the three worlds, see Figure 6.19.

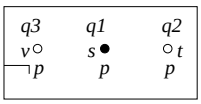

Figure 6.19: Update by a radical agent

In contrast to this, the conservative (or more suspicious) agent b would update his plausibility ordering in the manner depicted in Figure 6.20.

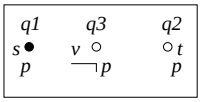

Figure 6.20: Update by conservative agents

We draw these orders separately here, though they will be part of one single total epistemic-doxastic update for the group when d's public announcement takes place. Next, the generally trusted source c tells agents a, b that q_2 is false. The above two models then change into the following new ones:

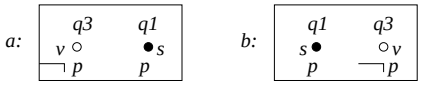

Figure 6.21: The final updates

We see from pictures that $B_a q_1$ and $B_b q_3$. Thus, a has acquired the right belief, but unfortunately, b has not! Thus, different revisors can get different convictions out of witnessing the same events, and indeed, some of them may be misled by correct information into believing false things! There is endless potential here for deceiving other agents – and even 'deception by the truth', which has already been observed by game theorists in the study of signaling games.[8]

Continuing with our example, dynamics of information flow is in principle *order-dependent*. What about the opposite order, where agent c speaks first, and only then the less-trusted d? Look at the original model again. After c's truthful announcement, both a and b will update their model into one with just the two possible worlds s and v of the last picture above. Then, when agent d tells them that p is true, the difference between the revision policies of a and b is immaterial: they can only raise the plausibility of p in one way, putting world s on top. Thus, both a and b acquire the right belief: as $B_a q_1$ and $B_b q_1$ hold.

To analyze this scenario in detail, one can use the machinery of Section 6.5 to express the types of agents qua revision policies (using the dynamic logics for belief revision discussed in Section 6.4), and then describe their interactions using a mixture of these type definitions and the general principles of dynamic epistemic-doxastic logic.

Admittedly, the preceding scenario is a bit contrived. More appealing scenarios of this sort would be variations of Muddy Children, where children revise beliefs rather than just updating knowledge, and where both skeptical and trusting children are around in the garden. In this way, belief revision policies would become concrete objects, whose workings can be determined precisely, and whose

[8]This phenomenon is also discussed in [BBS07] as a motivation for introducing a new epistemic attitude of 'safe belief', intermediate between belief and knowledge.

peculiarities may be exploited in more sophisticated puzzles of communication.

With these three scenarios, our discussion of interaction between diverse agents has come to an end. The main thrust of our investigation has been this. Once we have diverse agents inside one logical system, we can talk about the way they update their information and revise their beliefs. To deal with specific scenarios, we found that we needed the following additional ingredients: (a) more structured views of relevant events, (b) language extensions with types of agent, and their properties, where one may have to distinguish between the sender and the receiver of the information, and (c) mixtures of general dynamic-epistemic reasoning with specific information about agents. All this worked for information update, but we have also indicated how it applies to belief revision, when phrased in a dynamic logic format.

6.7 Conclusion and further challenges

This chapter has presented a more systematic discussion of different sources of diversity for rational agents than is usually found in the literature. More concretely, we showed how such diversity can be encoded in dynamic logics allowing for individual variation among agents. In particular, in the context of knowledge update, we made new proposals for modeling memory capacity, and defined a new version of product update for bounded k-memory agents. Next, in the context of belief revision, we showed how different revision policies can be put into one dynamic logic, allowing for great variation in revision and learning behavior.

Next, we pursued another essential phenomenon. Diversity among single agents is just a first ambition for logical modeling. But clearly, agents should not just 'live apart together'. Thus, we moved to the topic of interaction between agents of different types, discussing several scenarios which may arise then, having to do with different information processing, communication, and achieving of goals, when agents differ in their reliability, introspective powers, or belief revision policies. Our general conclusion was that these phenomena, too, can be modeled in our dynamic logics – but they need to be extended with explicit accounts of agents' types, and more structured informative events.

Even so, all this is only a beginning. There are several questions we would like to explore in the future. First, back to charting the sources of diversity, there remains the issue whether one can have a *general* view of the natural "parameters" that determine differences in behavior of logical agents. Our analysis does not provide such a general account, but at least, it shows more richness and uniformity than earlier ones. Second, even with all these parameters on the map, we have not yet found one framework for all these sources.

One particular area where this is true are agents' limitations in terms of inferential or computational powers. There is a body of work on the latter, witness

the survey chapter on 'Logic and Information' by [BM07] in the *Handbook of the Philosophy of Information*. In particular, the work of [Dun07], [Ågo04], and [Jag06] in computer science seems relevant here – as in the chapter by [Abr07] on the information content of computation in the same Handbook. Indeed, there are also long-standing connections with discussions of information content in the philosophical literature (cf. [Hin73]). The cited survey chapter discusses attempts at combining inferential diversity with observational and learning diversity as discussed in our chapter. Cf. [VQ07] for some further development.

Our next ambition would be to put all these features together in one plausible computational model of an agent as an information-processing and decision-making device, with modules for perception, memory, and inference which can communicate and share information.

Next, concerning interaction in diverse societies of agents, we have not yet looked at scenarios involving bounded memory – the way game theorists have when they discuss 'bounded rationality'. Here is where our dynamic epistemic or doxastic logics should meet up with current *game logics*, if we are to describe agents' longer-term strategies for collaboration, or competition, or more realistically, their frequent mixtures of both... Furthermore, with strategic behavior in the longer-term, our analysis of diversity in single update steps should meet up with temporal epistemic and doxastic logic, as explored in [FHMV95], [PR03], [BP06], and [Bon07].

Even so, we hope that our account of diversity and interaction is of use per se in placing the phenomenon on the map, while it also may provide a fresh look at current logical systems for information update and belief revision. Our cognitive and social reality is that different agents live together, and interact with each other, sometimes with remarkable success. This rich set of phenomena is not just a playground for psychologists or sociologists: it seems to be a legitimate challenge to logicians as well!

Chapter 7

Conclusions and Future Work

Conclusions

This dissertation has started from two issues concerning the functioning of rational agents that have been largely left aside since [Wri63]: *reasons* for preference, and *changes* in preference. Extrinsic reason-based preference was chosen as our main topic, and two models have been proposed for it in Chapters 2 and 3, respectively. Those models differ in their point of departure: object comparison versus priority order of propositions, but they have a common feature, in that preference and reasons come together.

In Chapter 2, I have shown how preference over propositions, derived from a primitive betterness relation over possible worlds, can be studied with techniques from dynamic epistemic logic. In particular, dynamic reduction axioms encode exactly how propositional preferences change when some new evaluative trigger such as a suggestion or command changes the betterness order. This brings a new methodology to traditional preference logic, while at the same time extending the scope of DEL.

In Chapter 3, preference over objects was studied in a fragment of first-order logic FOL. Preference over objects is now derived from a primitive base order of propositions in a priority sequence. This shows how logic can deal with basic ideas from Optimality Theory and related areas of 'optimal choice' in computer science and the social sciences. Moreover, I have shown how this, too, is compatible with DEL methodology, proposing dynamic operations on priority sequences with a complete set of reduction axioms.

In Chapter 4, a comparison between the two models of Chapters 2 and 3 showed that they are systematically related, and that they may fit together in various elegant mathematical ways. One example is a view of preference definition and preference change as related to more general preference merge between orderings coming from different sources. Another example is a grand two- or even three-level doxastic preferential predicate logic that can deal with the various notions of preference encountered in our intuitive daily reasoning.

In both Chapters 2 and 3, I have also shown how static preference representation and preference dynamics can live together with epistemic and doxastic structure, thereby doing justice to the intuitive entanglement of preference with knowledge and belief. I have brought these strands together in Chapter 4, showing how this all fits with a sequence of modal logics describing various 'degrees of entanglement'. This also allowed for further connections with belief revision theory, although we have by no means exhausted this analogy. Cf. the dissertation [Gir08] for a complementary agenda of logical themes at this rich interface.

In Chapters 5 and 6, I have then developed a logical perspective on much more general diversity of agents, trying to locate all aspects in which they can differ. Chapter 5 high-lights the dramatic difference between agents with perfect recall and agents with bounded memory. I have shown how both can be captured in dynamic epistemic logics, thereby dispelling the idea that DEL can only account for idealized agents, and making for connections with the theory of games with imperfect information. This diversity is then extended to logics defining policies for belief revision in both Chapters 5 and 6.

Thus I have proposed the basic ingredients for dynamic logics of agents with information update, belief revision, and preference upgrade. When adding these together in realistic settings, one must analyze the *interactions* of diverse agents. At the end of Chapter 6, I make a first step in this direction. I analyze a couple of scenarios in which different types of agents interact with each other, and propose a model for analyzing these.

Future work

This thesis proposes a rich model of diverse preference-driven rational agents based on dynamic logic. In doing so, many new questions have arisen, which have been noted along the way.

Some obvious open problems within the logical sphere are links between our various systems that are yet to be developed. First, we need to understand combined systems incorporating object relation transformers and constraint dynamics in greater generality, and Chapter 4 contained many leads in this direction. Next, the transition from Chapters 2, 3, 4 to the themes in Chapters 5, 6 suggests a more systematic merge of limitations on information dynamics (memory, inference, observation) and similar limitations on preference dynamics. To put it briefly, *how to model preference change for bounded agents?* We believe that our thesis supplies the right ingredients for doing so, but we have not done it yet.

Next, I have mostly considered preferences for single agents, while rational agency clearly involves *groups*. The current framework extends easily to interactive multi-agent systems in a purely formal manner. In Chapter 3 we made a start in investigating concepts of cooperation and competition by interpreting them in terms of reasons for preferences of the different agents. We also made a brief excursion on preference merge, and hence 'group preferences' in Chapter 4.

It should also be noted that preference change does not just involve myopic single steps. It often takes place in longer scenarios over *time*. To fully understand the temporal dynamics of preference, we need to integrate time into the current framework, as has been done for dynamic epistemic and doxastic logic in [BP06], [Bon07], and other publications.

Finally, going beyond the narrower ILLC world of logic and computation, there are also evident broader questions relating our present logical framework to other approaches.

In particular, I have adopted a *qualitative* approach to preference representation. But in areas like decision theory and social choice theory, usually, numerical utility functions represent preference. And likewise, for modeling beliefs under uncertainties, numerical probabilities are used widely. In the area of belief revision, and to some extent also DEL these days, this is a well-known interface. Can the logical systems for preference proposed in this thesis support *quantitative* utilities in a natural manner? I made a first attempt in [Liu06b], using DEL methodology to upgrade numerical 'plausibility values', using ideas from [Auc03], but much more remains to be done.

Moreover, this thesis has provided quite abstract models of reasons for preference and changes in preference. *Applying* these models to concrete scenarios in areas like decision theory and game theory seems a reasonable test for our proposals. For instance, in games, a player may have an initial preference over moves, but then, observing what her opponent plays may make her change her mind. In such a scenario, both preference and beliefs play a role, often at the same time (see [Ben06b]). Such considerations also affect how players compare propositions about the future course of a game. Thus, the usual solution procedure of Backward Induction involves a mixture of relative plausibility and preference between outcomes, as has been pointed out in [Ben02] and [DZ07]. Likewise, our models should be confronted with those in the philosophy of action, where preference supports rationality. The dissertation [Roy08] takes static preference logic and DEL-style dynamics to this arena in modeling information update and intentions, but it does not yet contain a full-fledged account of belief revision and preference change.

This concludes our summary of what this thesis has done, and what may be, and perhaps should be, done next.

Bibliography

[ABN98] H. Andréka, J. van Benthem, and I. Németi. Modal logics and bounded fragments of predicate logic. *Journal of Philosophical Logic*, 27:217–274, 1998.

[Abr96] S. Abramsky. Semantics of interaction: An introduction to game semantics. In P. Dybjer and A. Pitts, editors, *Proceedings CLiCS Summer School*, pages 1–31. Cambridge University Press, 1996.

[Abr07] S. Abramsky. Information, processes and games. In J. van Benthem and P. Adriaans, editors, *Handbook of Philosophy of Information*. 2007. To appear.

[AGM85] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[Ågo04] T. Ågotnes. *A Logic of Finite Syntactic Epistemic States*. PhD thesis, Department of Informatics, University of Bergen, 2004.

[AJL06] N. Alechina, M. Jago, and B. Logan. Modal logics for communicating rule-based agents. In A. Perini G. Brewka, S. Coradeschi and P. Traverso, editors, *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI 2006)*. IOS Press, 2006.

[AJL07] N. Alechina, M. Jago, and B. Logan. Belief revision for rule-based agents. In J. van Benthem, S. Ju, and F. Veltman, editors, *A Meeting of the Minds–Proceedings of the Workshop on Logic, Rationality and Interaction*, pages 99–112. King's College Publications, 2007.

[ALW04] N. Alechina, B. Logan, and M. Whitsey. A complete and decidable logic for resource-bounded agents. In *AAMAS 2004*, pages 606–613, 2004.

[Åqv87] L. Åqvist. *Introduction to Deontic Logic and the Theory of Normative Systems*. Naples: Bibliopolis, 1987.

[Åqv94] L. Åqvist. Deontic logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2. Dordrecht: Kluwer, 1994.

[ARS02] H. Andréka, M. Ryan, and P-Y. Schobbens. Operators and laws for combining preferential relations. *Journal of Logic and Computation*, 12:12–53, 2002.

[Auc03] G. Aucher. A combined system for update logic and belief revision. Master's thesis, MoL-2003-03. ILLC, University of Amsterdam, 2003.

[Auc05] G. Aucher. How our beliefs contribute to interpret actions. In M. Pechoucek, P. Petta, and L.Z. Varga, editors, *CEEMAS 2005*, pages 276–286. Springer, 2005. LNAI 3690.

[Axe84] R. Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[BB07] P. Blackburn and J. van Benthem. Modal logic: A semantic perspective. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, chapter 1, pages 1–85. Elsevier, 2007.

[BBS07] A. Baltag, J. van Benthem, and S. Smets. A dynamic-logical approach to interactive epistemology. Working paper, ILLC, University of Amsterdam, 2007.

[BC03] J. van Benthem and B. ten Cate. Automata and update agents in event trees. Working paper, Department of Philosophy, Stanford University, 2003.

[BCD+93] S. Benferhat, C. Cayol, D. Dubois, J. Lang, and H. Prade. Inconsistency management and prioritized syntax-based entailment. In *Proceedings of IJCAI'93*, pages 640–645, 1993.

[BE07] D. Bonnay and P. Egré. A non-standard semantics for inexact knowledge with introspection. Working paper, IHPST, Paris, 2007.

[BEF93] J. van Benthem, J. van Eijck, and A. Frolova. Changing preferences. Technical Report, CS-93-10, Centre for Mathematics and Computer Science, Amsterdam, 1993.

[BEK06] J. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204:1620–1662, 2006.

[Ben82] J. van Benthem. Later than late: On the logical origin of the temporal order. *Pacific Philosophical Quarterly*, 63:193–203, 1982.

[Ben96] J. van Benthem. *Exploring Logical Dynamics*. CSLI Publication, Stanford, 1996.

[Ben99] J. van Benthem. Modal correspondence theory. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 3, pages 325–408. Dordrecht: Kluwer, second edition, 1999. Reprint with addenda.

[Ben00] J. van Benthem. Information transfer across Chu spaces. *Logic Journal of the IGPL*, 8:719–731, 2000.

[Ben01] J. van Benthem. Games in dynamic-epistemic logic. *Bulletin of Economic Research*, 53:219–248, 2001.

[Ben02] J. van Benthem. Extensive games as process models. *Journal of Logic, Language and Information*, 11:289–313, 2002.

[Ben03] J. van Benthem. Logic games are complete for game logics. *Studia Logica*, 75:183–203, 2003.

[Ben04a] J. van Benthem. Local versus global update in games. Working paper, Department of Philosophy, Stanford University, 2004.

[Ben04b] J. van Benthem. A mini-guide to logic in action. In *Philosophical Researches: Special Issue in Logic*, pages 21–30. The Chinese Association of Logic, Beijing, 2004.

[Ben06a] J. van Benthem. 'One is a lonely number': On the logic of communication. In P. Koepke Z. Chatzidakis and W. Pohlers, editors, *Logic Colloquium, ASL Lecture Notes in Logic 27*. AMS Publications, Providence (R.I.), 2006. Technical Report, PP-2002-27, ILLC, University of Amsterdam.

[Ben06b] J. van Benthem. Rationalizations and promises in games. In *Philosophical Trend: Special Issue in Logic*, pages 1–6. The Chinese Association of Logic, Beijing, 2006.

[Ben07a] J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logic*, 17:129–156, 2007. Technical Report, PP-2006-11, ILLC, University of Amsterdam.

[Ben07b] J. van Benthem. Priority update and merging preorders. Working paper. ILLC, University of Amsterdam, 2007.

[Ben07c] J. van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9:13–45, 2007.

[Ben08] J. van Benthem. Rational animals, a logical jungle? To appear in the *Journal of Peking University* (Philosophy and Social Sciences), 2008.

[BG07] T. Brauner and S. Ghilardi. First-order modal logic. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, pages 549–620. Elsevier, 2007.

[BGK06] J. van Benthem, J. Gerbrandy, and B. Kooi. Dynamic update with probabilities. Technical Report, PP-2006-21, ILLC, University of Amsterdam, 2006.

[BGP07] J. van Benthem, J. Gerbrandy, and E. Pacuit. Merging frameworks for interaction: DEL and ETL. In D. Samet, editor, *Proceedings of TARK*. 2007.

[BHT06] J. Broersen, A. Herzig, and N. Troquard. Embedding alternating-time temporal logic in strategic STIT logic of agency. *Journal of Logic and Computation*, 16:559–578, 2006.

[BL04] J. van Benthem and F. Liu. Diversity of logical agents in games. *Philosophia Scientiae*, 8:163–178, 2004.

[BL07] J. van Benthem and F. Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logic*, 17:157–182, 2007. Technical Report, PP-2005-29, ILLC, University of Amsterdam.

[Bla93] P. Blackburn. Nominal tense logic. *Notre Dame Journal of formal logic*, 34:65–83, 1993.

[BM07] J. van Benthem and M. Martínez. The stories of logic and information. In J. van Benthem and P. Adriaans, editors, *Handbook of Philosophy of Information*. 2007. To appear.

[BMS98] A. Baltag, L.S. Moss, and S. Solecki. The logic of common knowledge, public announcements, and private suspicions. In I. Gilboa, editor, *Proceedings of the 7th conference on theoretical aspects of rationality and knowledge (TARK 98)*, pages 43–56, 1998.

[Bol83] F. Bolle. On Sen's second-order preferences, morals, and decision theory. *Erkenntnis*, 20:195–205, 1983.

[Bon04] G. Bonanno. Memory and perfect recall in extensive games. *Games and Economic Behavior*, 47:237–256, 2004.

[Bon07] G. Bonanno. Belief revision in a temporal framework. Working paper, University of California at Davis, 2007.

[BOR06] J. van Benthem, S. van Otterloo, and O. Roy. Preference logic, conditionals and solution concepts in games. In H. Lagerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, pages 61–77. Uppsala Philosophical Studies 53, 2006.

[Bou93] C. Boutilier. A modal characterization of defeasible deontic conditionals and conditional goals. In *AAAI Spring Symposium on Reasoning about Mental States*, pages 30–39. 1993.

[Bou94] C. Boutilier. Conditional logics of normality: A modal approach. *Artificial Intelligence*, 68:87–154, 1994.

[BP06] J. van Benthem and E. Pacuit. The tree of knowledge in action: Towards a common perspective. In G. Governatori, I. Hodkinson, and Y. Venema, editors, *Proceedings of Advances in Modal Logic (AiML 2006)*. Uppsala Philosophical Studies 53, 2006.

[BPX01] N. Belnap, M. Perloff, and M. Xu. *Facing the Future*. Oxford University Press, Oxford, 2001.

[BRG07] J. van Benthem, O. Roy, and P. Girard. Everything else being equal: A modal logic approach to ceteris paribus preferences. Technical Report, PP-2007-09, ILLC, University of Amsterdam, 2007.

[BRP01] S. Barbera, W. Rossert, and Prasanta K. Pattanaik. Ranking sets of objects. Departement de sciences economiques, Universite de Montreal. See http://www.sceco.umontreal.ca/publications/etext/2001-02.pdf, 2001.

[Bru04] B. de Bruin. *Explaining Games: On the Logic of Game Theoretic Explanations*. PhD thesis, ILLC, University of Amsterdam, 2004.

[BRV01] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2001.

[BS06a] A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. In *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory (LOFT 06)*, Liverpool, 2006.

[BS06b] A. Baltag and S. Smets. The logic of conditional doxastic actions: A theory of dynamic multi-agent belief revision. In S. Artemov and R. Parikh, editors, *Proceedings of the Workshop on Rationality and Knowledge*, ESSLLI, Malaga, 2006.

[BS08] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. To appear in G. Bonanno, W. van der Hoek, M. Wooldridge, editors, *Texts in Logic and Games*, 2008.

[CEL06] Y. Chevaleyre, U. Endriss, and J. Lang. Expressive power of weighted propositional formulas for cardinal preference modelling. In *Proceedings of KR 2006*, pages 145–152. AAAI Press, 2006.

[CL90] P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–361, 1990.

[CMLLM04] S. Coste-Marquis, J. Lang, P. Liberatore, and P. Marquis. Expressive power and succinctness of propositional languages for preference representation. In *Proceedings of KR 2004*. AAAI Press, 2004.

[Cre71] M.J. Cresswell. A semantics for a logic of 'better'. *Logique et Analyse*, 14:775–782, 1971.

[CS66a] R.M. Chisholm and E. Sosa. Intrinsic preferability and the problem of supererogation. *Synthese*, 16:321–331, 1966.

[CS66b] R.M. Chisholm and E. Sosa. On the logic of 'intrinsically better'. *American Philosophical Quarterly*, 3:244–249, 1966.

[Dég07a] C. Dégremont. Beliefs and expectations in time. Working paper, ILLC, University of Amsterdam, 2007.

[Dég07b] C. Dégremont. Epistemic foundations of extensive games and qualitative belief revision. Working paper, ILLC, University of Amsterdam, 2007.

[Dem05] S. Demri. A reduction from DLP to PDL. *Journal of Logic and Computation*, 15:767–785, 2005.

[DHK07] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Berlin: Springer, 2007.

[Dit05] H. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese(Knowledge, Rationality & Action)*, 147:229–275, 2005.

[DSW91] J. Doyle, Y. Shoham, and M.P. Wellman. A logic of relative desire. In *Proceedings of 6th International Symposium on Methodologies for Intelligence Systems*, pages 16–31, 1991.

[DT99] J. Doyle and R.H. Thomason. Background to qualitative decision theory. *AI Magazine*, 20:55–68, 1999.

[Dun95] P.H. Dung. An argumentation-theoretic foundation for logic programming. *Journal of Logic Programming*, 22:151–177, 1995.

[Dun07] J. M. Dunn. Information in computer science. In J. van Benthem and P. Adriaans, editors, *Handbook of Philosophy of Information*. 2007. To appear.

[DW94] J. Doyle and M.P. Wellman. Representing preferences as Ceteris Paribus comparatives. *Working Notes of the AAAL Symposium on Decision-Theoretic Planning*, 1994.

[DW07] F. Dechesne and Y. Wang. Dynamic epistemic verification of security protocols: Framework and case study. In J. van Benthem, S. Ju, and F. Veltman, editors, *A Meeting of the Minds–Proceedings of the Workshop on Logic, Rationality and Interaction*, pages 129–144. King's College Publications, 2007.

[DZ07] C. Dégremont and J. Zvesper. Dynamic logic for cognitive actions in extensive games. Working paper, ILLC, University of Amsterdam, 2007.

[EC92] J. van Eijck and G. Cepparello. Dynamic modal predicate logic. Technical Report CS-R9237, 1992.

[Eck82] J. van Eck. *A system of temporally relative modal and deontic predicate logic and its philosophical applications*. PhD thesis, 1982.

[Egr04] P. Egré. *Propositional Attitudes and Epistemic Paradoxes*. PhD thesis, Universite Paris 1 et IHPST, 2004.

[FH85] R. Fagin and J.Y. Halpern. Belief, awareness, and limited reasoning. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 480–490, 1985.

[FHMV95] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. The MIT Press, 1995.

[Fis73] P.C. Fishburn. *The Theory of Social Choice*. Princeton University Press, 1973.

[Fis99] P.C. Fishburn. Preference structures and their numerical representations. *Theoretical Computer Science*, 217:359–383, 1999.

[FM98] M. Fitting and R.L. Mendelsohn. *First-order Modal Logic*. Dordrecht: Kluwer, 1998.

[Ger99] J. Gerbrandy. *Bisimulation on Planet Kripke*. PhD thesis, ILLC, University of Amsterdam, 1999.

[Gir08] P. Girard. *Modal logics for belief and preference change*. PhD thesis, Stanford University, 2008. To appear.

[GM88] P. Gärdenfors and D. Makinson. Revisions of knowledge systems using epistemic entrenchment. In *Proceedings of TARK*, pages 83–95, 1988.

[Gol05] E.B. Goldstein. *Cognitive Psychology - Connecting, Mind Research, and Everyday Experience*. Thomson Wadsworth, 2005.

[GR95] P. Gärdenfors and H. Rott. Belief revision. In D.M. Gabbay, C.J. Hogger, and J.A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 4. Oxford University Press, 1995.

[Gro88] A. Grove. Two modelings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[Gro91] B.N. Grosof. Generalising prioritization. In *Proceedings of KR 91*, pages 289–300. Morgan Kaufmann, 1991.

[GS98] D. Gabbay and V. Shehtman. Products of modal logics. *Logic Journal of the IGPL*, 6:73–146, 1998. Part 1.

[GTtARG99] G. Gigerenzer, P. Todd, and the ABC Research Group. *Simple Heuristics that Make us Smart*. New York: Oxford University Press, 1999.

[Hal57] S. Halldén. *On the Logic of "Better"*. Lund, 1957.

[Hal97] J.Y. Halpern. Defining relative likelihood in partially-ordered preferential structure. *Journal of Artificial Intelligence Research*, 7:1–24, 1997.

[Han68] B. Hansson. Fundamental axioms for preference relations. *Synthese*, 18:423–442, 1968.

[Han90a] S.O. Hansson. Defining 'good' and 'bad' in terms of 'better'. *Notre Dame of Journal of Formal Logic*, 31:136–149, 1990.

[Han90b] S.O. Hansson. Preference-based deontic logic. *Journal of Philosophical Logic*, 19:75–93, 1990.

[Han95] S.O. Hansson. Changes in preference. *Theory and Decision*, 38:1–28, 1995.

[Han01a] S.O. Hansson. Preference logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 4, chapter 4, pages 319–393. Dordrecht: Kluwer, 2001.

[Han01b] S.O. Hansson. *The Structure of Values and Norms*. Cambridge University Press, 2001.

[Har04] P. Harrenstein. *Logic in Conflict. Logical Explorations in Strategic Equilibrium*. PhD thesis, Utrecht University, 2004.

[HC96] G.E. Hughes and M.J. Cresswell. *A New Introduction to Modal Logic*. Routledge: London and New York, 1996.

[Hen01] V.F. Hendricks. *The Convergence of Scientific Knowledge - a view from the limit*. Studia Logica Library Series: Trends in Logic. Kluwer Academic Publishers, 2001.

[Hen03] V.F. Hendricks. Active agents. *Journal of Logic, Language and Information*, 12:469–495, 2003.

[HGY06] S.O. Hansson and T. Grüne-Yanoff. Preferences. In *Stanford Encyclopedia of Philosophy*. Stanford, 2006. http://plato.stanford.edu/entries/preferences/.

[HHB07] H. Hodges, W. Hodges, and J. van Benthem, editors. *Logic and Psychology*. 2007. Guest issue of Topoi, to appear.

[Hin73] J. Hintikka. *Logic, Language-Games and Information*. Clarendon Press, Oxford, 1973.

[HJT90] W. van der Hoek, J. Jaspars, and E. Thijsse. A general approach to multi-agent minimal knowledge. In M. Ojeda-Aciego, I.P. Guzman, G. Brewka, and L.M. Pereira, editors, *Proceedings of the 7th European Workshop on Logics in Artificial Intelligence (JELIA 2000)*, pages 254–268. Springer-Verlag, Heidelberg, 1990. LNAI 1919.

[HK02] D. Houser and R. Kurzban. Revealed preference, belief, and game theory. *Economics and Philosophy*, 16:99–115, 2002.

[HKT00] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. The MIT Press, 2000.

[HLP00] A. Herzig, J. Lang, and T. Polacsek. A modal logic for epistemic tests. In *Proceedings of the ECAI 2000*, Berlin, 2000.

[Höt03] T. Hötte. A model for epistemic games. Master's thesis, MoL-2003-05. ILLC, University of Amsterdam, 2003.

[HT06] A. Herzig and N. Troquard. Knowing how to play: uniform choices in logics of agency. In *AAMAS 2006*, pages 209–216, 2006.

[Hug80] R. Hughes. Rationality and intransitive preferences. *Analysis*, 40:132–134, 1980.

[HW03] W. van der Hoek and M. Wooldridge. Towards a logic of rational agency. *Logic Journal of the IGPL*, 11:133–157, 2003.

[Jag06] M. Jago. *Logics for Resource-Bounded Agents*. PhD thesis, University of Nottingham, 2006.

[Jef65] R.C. Jeffrey. *The Logic of Decision*. Chicago: University of Chicago Press, 1965.

[Jen67] R.E. Jennings. Preference and choice as logical correlates. *Mind*, 76:556–567, 1967.

[JL06] D. de Jongh and F. Liu. Optimality, belief and preference. In S. Artemov and R. Parikh, editors, *Proceedings of the Workshop on Rationality and Knowledge*. ESSLLI, Malaga, 2006. Technical Report, PP-2006-38, ILLC, University of Amsterdam.

[Kat07] S. Katsuhiko. A hybridization of irreflexive modal logics. *Electronic Notes in Theoretical Computer Science*, 174:95–111, 2007.

[Kon88] K. Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35:343–382, 1988.

[Koo07] B. Kooi. Dynamic term-modal logic. In J. van Benthem, S. Ju, and F. Veltman, editors, *A Meeting of the Minds–Proceedings of the Workshop on Logic, Rationality and Interaction*, pages 173–186. King's College Publications, 2007.

[Kra81] A. Kratzer. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, 10:201–216, 1981.

[KZ03] A. Kurucz and M. Zakharyaschev. A note on relativised products of modal logics. In P. Balbiani, N.-Y. Suzuki, F. Wolter, and M. Zakharyaschev, editors, *Advances in Modal Logic*, volume 4, pages 221–242. King's College Publications, 2003.

[Lee84] R. Lee. Preference and transitivity. *Analysis*, 44:129–134, 1984.

[Leh95] D. Lehmann. Another perspective on default reasoning. *Annals of Mathematics and Artificial Intelligence*, 15:61–82, 1995.

[Len83] W. Lenzen. On the representation of classificatory value structures. *Theory and Decision*, 15:349–369, 1983.

[Lev90] H. J. Levesque. All I know: A study in autoepistemic logic. *Artificial Intelligence Journal*, 42:381–386, 1990.

[Lew73] D. Lewis. *Counterfactuals*. Oxford: Blackwell, 1973.

[Lew81] D. Lewis. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic*, 10:217–234, 1981.

[LHM96] B. van Linder, W. van der Hoek, and J-J.Ch. Meyer. Formalising motivational attitudes of agents: On preferences, goals and commitments. In M. Wooldridge, J. Mueller, and M. Tambe, editors, *Intelligent Agents Volume II - Agent Theories, Architectures, and languages (ATAL '96)*, pages 17–32. Berlin: Springer, 1996.

[Liu04] F. Liu. Dynamic variations: Update and revision for diverse agents. Master's thesis, MoL-2004-05. ILLC, University of Amsterdam, 2004.

[Liu06a] F. Liu. Diversity of agents. In *Proceedings of the Workshop on Logics for Resource Bounded Agents*, ESSLLI, Malaga, 2006. Technical Report, PP-2006-37, ILLC, University of Amsterdam.

[Liu06b] F. Liu. Preference change and information processing. In *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory (LOFT 06)*, Liverpool, 2006. Technical Report, PP-2006-41, ILLC, University of Amsterdam.

[Liu07] F. Liu. Diversity of agents and their interaction. To appear in *Journal of Logic, Language and Information*. Technical Report, PP-2007-01, ILLC, University of Amsterdam, 2007.

[LTW03] J. Lang, L. van der Torre, and E. Weydert. Hidden uncertainty in the logical representation of desires. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, 2003.

[Mey88] J-J.Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29:109–136, 1988.

[Mey96] R. van der Meyden. The dynamic logic of permission. *Journal of Logic and Computation*, 6:465–479, 1996.

[Moo85] R. Moore. A formal theory of knowledge and action. In J.R. Hobbs and R.C. Moore, editors, *Formal Theories of the Common Sense World*. Ablex Publishing, Norwood, NJ, 1985.

[Nay94] A. Nayak. Iterated belief change based on epistemic entrenchment. *Erkentnis*, 41:353–390, 1994.

[OR94] M. Osborne and A. Rubinstein. *A Course in Game Theory*. The MIT Press, Cambridge (Mass.), 1994.

[Ott05] S. van Otterloo. *A Strategic Analysis of Multi-agent Protocols*. PhD thesis, Liverpool University, UK, 2005.

[Pla89] J.A. Plaza. Logics of public announcements. In *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, 1989.

[PPC06] E. Pacuit, R. Parikh, and E. Cogan. The logic of knowledge based on obligation. *Knowledge, Rationality and Action*, 149:311–341, 2006.

[PR03] R. Parikh and R. Ramanujam. A knowledge-based semantics of messages. *Journal of Logic, Language and Information*, 12:453–467, 2003.

[PS93] A. Prince and P. Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden, Ma: Blackwell, 1993.

[PW04] R. Pucella and V. Weissmann. Reasoning about dynamic policies. In *Proceedings FoSSaCS-7*, Lecture Notes in Computer Science 2987, pages 453–467, 2004.

[Res66] N. Rescher. Notes on preference, utility, and cost. *Synthese*, 16:332–343, 1966.

[RG91] A.S. Rao and M.P. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1991.

[Rod01] B. Rodenhäuser. Updating epistemic uncertainty: An essay in the logic of information change. Master's thesis, MoL-2001-07. ILLC, University of Amsterdam, 2001.

[Roh05] Ph. Rohde. *On Games and Logics over Dynamically Changing Structures*. PhD thesis, Department of Informatics, Technische Hochschule Aachen (RWTH), 2005.

[Roo07] R. van Rooij. Semi-orders and satisficing behavior. Working paper, ILLC, University of Amsterdam, 2007.

[Rot01] H. Rott. *Change, Choice and Inference: A Study of Belief and Revision and Nonmonotonic Reasoning*. New York: Oxford University Press, 2001.

[Rot03] H. Rott. Basic entrenchment. *Studia Logica*, 73:257–280, 2003.

[Rot06] H. Rott. Shifting priorities: Simple representations for 27 iterated theory change operators. In H. Langerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, pages 359–384. Uppsala Philosophical Studies 53, 2006.

[Rot07] H. Rott. Information structures in belief revision. In J. van Benthem and P. Adriaans, editors, *Handbook of Philosophy of Information*. 2007. To appear.

[Roy08] O. Roy. *Thinking before Acting: Intentions, Logic and Rational Choice*. PhD thesis, ILLC, University of Amsterdam, 2008. To appear.

[Rub98] A. Rubinstein. *Modeling Bounded Rationality*. The MIT Press, 1998.

[Sav54] L.J. Savage. *The Foundations of Statistics*. Wiley, New York, 1954.

[Sch75] G. Schumm. Remark on a logic of preference. *Notre Dame Journal of Formal Logic*, 16:509–510, 1975.

[Sch97] S. Scheffler. Relationships and responsibilities. *Philosophy and Public Affairs*, 26:189–209, 1997.

[Sen71] A. Sen. Choice functions and revealed preference. *Review of Economic Studies*, 38:307–317, 1971.

[Sen73] A. Sen. Behaviour and the concept of preference. *Economica*, 40:241–259, 1973.

[Ser04] M. Sergot. $(C+)^{++}$: An action language for modelling norms and institutions. Technical Report 8, Department of Computing, Imperial College, London, 2004.

[Sev06] M. Sevenster. *Branches of Imperfect Information: Logic, Games, and Computation*. PhD thesis, ILLC, University of Amsterdam, 2006.

[SGG03] A. Silberschatz, P.B. Galvin, and G. Gagne. *Operating System Concepts*. John Wiley & Sons, Inc., New York, NY, USA, 2003.

[Sho88] Y. Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*. The MIT Press: Cambridge, MA, 1988.

[Smo04] P. Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Publishers, 2004.

[Sny04] J. Snyder. Product update for agents with bounded memory. Manuscript, Department of Philosophy, Stanford University, 2004.

[Spo88] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change and Statistics II*, pages 105–134. Kluwer, Dordrecht, 1988.

[Tho00] R. Thomason. Desires and defaults: A framework for planning with inferred goals. In *Proceedings of KR 2000*, pages 702–713, 2000.

[Tor97] L. van der Torre. *Reasoning about Obligations: Defeasibility in Preference-based Deontic Logic*. PhD thesis, Rotterdam, 1997.

[Tra85] R.W. Trapp. Utility theory and preference logic. *Erkenntnis*, 22:301–339, 1985.

[TT98] L. van der Torre and Y. Tan. The temporal analysis of chisholm's paradox. In *AAAI '98/IAAI '98: Proceedings of the 15th national/10th conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 650–655, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence.

[TT99] L. van der Torre and Y. Tan. An update semantics for deontic reasoning. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems*, pages 73–90. IOS Press, 1999.

[Tve69] A. Tversky. Intransitivity of preferences. *Psychological Review*, 76:31–48, 1969.

[Vel96] F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261, 1996.

[VQ07] F.R. Velazquez-Quesada. Relating abduction and belief revision. Working paper, ILLC, University of Amsterdam, 2007.

[Was00] R. Wassermann. *Resource Bounded Belief Revision*. PhD thesis, ILLC, University of Amsterdam, 2000.

[Wed03] R. Wedgwood. Choosing rationally and choosing correctly. In S. Stroud and C. Tappolet, editors, *Weakness of Will and Practical Irrationality*, pages 201–229. Oxford University Press, 2003.

[Woo00] M. Wooldridge. *Reasoning about Rational Agents*. The MIT Press (Cambridge, Massachussetts/London, England), 2000.

[Wri51] G.H. von Wright. Deontic logic. *Mind*, 60:1–15, 1951.

[Wri63] G.H. von Wright. *The Logic of Preference*. Edinburgh, 1963.

[Wri72] G.H. von Wright. The logic of preference reconsidered. *Theory and Decision*, 3:140–169, 1972.

[Yam06] T. Yamada. Acts of commands and changing obligations. In K. Inoue, K. Satoh, and F. Toni, editors, *Proceedings of the 7th Workshop on Computational Logic in Multi-Agent Systems (CLIMA VII)*, 2006. Revised version appeared in LNAI 4371, pages 1-19, Springer-Verlag, 2007.

[Yam07] T. Yamada. Logical dynamics of some speech acts that affect obligations and preferences. In J. van Benthem, S. Ju, and F. Veltman, editors, *A Meeting of the Minds–Proceedings of the Workshop on Logic, Rationality and Interaction*, pages 275–289. King's College Publications, 2007.

[Yap06] A. Yap. Product update and looking backward. Technical Report, PP-2006-39, ILLC, University of Amsterdam, 2006.

[Zar03] B. Zarnic. Imperative change and obligation to do. In K. Segerberg and R. Sliwinski, editors, *Logic, Law, Morality: Thirteen Essays in Practical Philosophy in Honour of Lennart Aqvist*, pages 79–95. Uppsala philosophical studies 51. Uppsala: Department of Philosophy, Uppsala University, 2003.

[Zve07] J. Zvesper. How to keep on changing your mind. In J. van Benthem, S. Ju, and F. Veltman, editors, *A Meeting of the Minds–Proceedings of the Workshop on Logic, Rationality and Interaction*, pages 291–308. King's College Publications, 2007.

Index

Summary

This thesis investigates two main issues concerning the behavior of rational agents, preference dynamics and agent diversity.

We take up two questions left aside by von Wright, and later also the multitude of his successors, in his seminal book *Logic of Preference* in 1963: *reasons* for preference, and *changes* in preference. Various notions of preference are discussed, compared and further correlated in the thesis. In particular, we concentrate on extrinsic preference. Contrary to intrinsic preference, extrinsic preference is reason-based, i.e. one's preference for one option over another has a reason. A logical model is proposed and its properties are determined. Dynamics come in naturally, since reasons for the preferences can change. Logical systems and formal results regarding dynamical preference change are then presented.

Preference arises from comparisons between alternatives. A first option is to compare situations. Abstractly speaking, preferences are in this case between propositions, viewed as sets of possible worlds. The reasons can then be based on a 'betterness' relation over possible worlds. Propositional preference arises as a lift from this primitive relation. A standard modal logical approach is taken and we use a modality for the betterness relation in the language. We then model preference change by techniques from dynamic epistemic logic (DEL), where a typical action, e.g. a suggestion or a command can change the betterness ordering of the worlds, and thereby the propositional preference. Dynamic reduction axioms are obtained to encode exactly how such a change takes place. We obtain a complete dynamic preference logic.

A second option is to compare objects as such. Concretely, properties of the objects often determine the preference over the objects. Properties are now the reasons. Inspired by Optimality Theory (OT), we propose a *priority sequence*, an ordering of properties. Various ways of getting a preference from the priority sequence are investigated, though we mostly follow the OT approach. We use a fragment of first-order logic to describe the situation. Here, on the dynamic side, it is priority change that leads to preference change. Using the DEL methodology

again we propose a complete set of reduction axioms concerning the possible dynamic operations on priority sequences.

Not surprisingly, the above two views are closely related. After all, possible worlds can be thought of as objects! On the basis of a systematic comparison of the two views, we develop a *two level perspective*, in which the models themselves are structured in layers. In particular, correspondence results between the changes at the level of the possible worlds and the changes at the level of the priority sequences are proved. We end up by sketching a two-level preferential predicate logic to describe more complex circumstances in which situations and objects are compared simultaneously.

But we do not see this as the whole story. Preference does not live by itself, it is often intermingled with epistemic notions of knowledge and belief. One can have different intuitions about how this entanglement operates. A few options are discussed, and proposals for logical models are presented. When moving to dynamics, we now see a picture of knowledge update, belief revision and preference change taking place symbiotically, often unconsciously as in real life.

The resulting picture in the thesis is one of agents that process information and adjust beliefs and preferences in many different ways. There is no logically prescribed unique norm for doing this. The second part of the thesis takes this general phenomenon of diversity of agents as its focus, since it raises many issues for logical systems and the idealized agents which they normally presuppose. In reality, agents can differ across a wide spectrum of cognitive abilities and habits: in their memory capacity, observation power, inferential power, introspective ability, and revision policies when facing new information.

Two kinds of agents, perfect recall agents, and memory-free agents are studied thoroughly, in the setting of playing games. We show how current dynamic logics can be 'parametrized' to allow for this memory diversity, with new characterizations of agent types resulting in complete dynamic-epistemic logics. The other dimension of the reality of diverse agents is that, however different they are, they often do manage to coordinate with each other successfully. While this theme has been prominent in game theory and multi-agent systems, it has received hardly any attention in logic. We analyze this interaction between different agents by looking at concrete scenarios which model their types explicitly.

Finally, we analyze the issue of agent diversity in its generality, discussing what dynamic logics would have to look like to become a full-fledged account of agents of different capacities and tendencies that pursue and sometimes achieve their goals in irreducibly social settings.

Samenvatting

Dit proefschrift onderzoekt twee centrale vragen inzake het gedrag van rationele actoren, dynamische verandering in hun voorkeuren, en diversiteit in vermogens.

We gaan in op twee belangrijke thema's die destijds terzijde werden geschoven door von Wright, en later zijn vele volgelingen, in zijn baanbrekende boek *Logic of Preference* uit 1963. Die thema's zijn de *redenen* die mensen hebben vóór hun voorkeuren, en de *veranderingen* in hun voorkeuren. Verschillende begrippen van voorkeur worden besproken, vergeleken, en systematisch met elkaar in verband gebracht in dit proefschrift. In het bijzonder richten we ons op extrinsieke voorkeuren. In tegenstelling tot intrinsieke voorkeuren zijn extrinsieke gebaseerd op redenen, onze voorkeur voor het één boven het ander heeft een reden. We ontwikkelen een logisch model en de eigenschappen ervan worden vastgesteld. Dynamiek komt hierbij op een natuurlijke wijze aan de orde, omdat redenen voor voorkeuren kunnen veranderen. Logische systemen en formele resultaten inzake dynamische verandering van voorkeuren worden gepresenteerd.

Voorkeuren ontstaan door vergelijken van alternatieven. Een eerste aanpak gebruikt situaties. Abstract gezien lopen voorkeuren in dit geval tussen proposities opgevat als verzamelingen mogelijke werelden. Redenen kunen dan worden gebaseerd op een primitieve vergelijkingsrelatie van 'beter' tussen mogelijke werelden. Propositionele voorkeuren ontstaan nu als een overdracht vanuit deze primitieve relatie. We volgen een standaard modale aanpak met een modaliteit voor de 'beter' relatie in de taal. Vervolgens modelleren we verandering van voorkeur met technieken uit de dynamisch-epistemische logica (DEL), waar een karakteristieke handeling, bijvoorbeeld een suggestie of een bevel, de relatie 'beter' tussen werelden kan veranderen, en daarmee ook de voorkeur tussen proposities. Dynamische reductie-axioma's leggen dan precies vast hoe zo'n verandering plaats vindt. Aldus ontstaat een volledige dynamische preferentielogica.

Een tweede aanpak werkt door objecten als zodanig te vergelijken. In concreto worden voorkeuren tussen objecten vaak bepaald door eigenschappen van die objecten. Die eigenschappen zijn dan onze 'redenen'. Geïnspireerd door

de taalkundige Optimaliteitstheorie (OT), stellen we een notie voor van *'prioriteitsreeks'*, een ordening van eigenschappen. Verschillende manieren worden onderzocht om voorkeuren af te leiden uit een prioriteitsreeks, hoewel we doorgaans de OT-manier volgen. We gebruiken een fragment van de eerste-orde logica om de situatie te beschrijven. Dynamisch gezien is het hier verandering van prioriteiten die leidt tot verandering van voorkeuren. Weer gebruikmakend van de DEL-methodiek stellen we een volledig stel reductie-axioma's voor die passen bij de mogelijke dynamische operaties op proriteitsreeksen.

Uiteraard staan de twee gezichtspunten tot nu toe met elkaar in verband. Zo kunnen bijvoorbeeld mogelijke werelden zelf als objecten worden gezien! Daartoe ontwikkelen we een *twee-niveau perspectief*. Modellen zijn nu zelf gestructureerd, en we bewijzen correspondentieresultaten tussen veranderingen op het niveau der mogelijke werelden en der prioriteitsreeksen. Tenslotte schetsen we een preferentiële predikaatlogica met twee niveaus, die meer complexe scenarios kan beschrijven waarin situaties en objecten tegelijkertijd worden vergeleken.

Maar dit is nog niet het volledige verhaal. Voorkeuren leven niet op zich, maar zijn doorgaans verstrengeld met epistemische begrippen van kennis en geloof. Er bestaan verschillende intuïties over hoe deze samenhang precies werkt. Enkele opties worden besproken, inclusief hun logische modellering. Aan de dynamische kant ontstaat dan een beeld van simultane kennis-aanpassing, geloofsherziening, en verandering in voorkeuren, vaak onbewust, net als in het dagelijks leven.

Het algemene perspectief dat hiermee in dit proefschrift naar voren komt is er een van actoren die gestaag informatie verwerken, en hun meningen en voorkeuren daarbij op allerlei manieren aanpassen. Maar er is niet één unieke logische regel die zegt hoe dit dient te gebeuren. Het tweede deel van dit proefschrift stelt daarom het algemene verschijnsel van diversiteit van actoren centraal, omdat dit vele nieuwe vragen opwerpt voor bestaande logische systemen en de geïdealiseerde actoren die daarin doorgaans worden gepostuleerd. In werkelijkheid kunnen actoren immers verschillen in een breed spectrum van cognitieve vermogens en gewoontes: geheugencapaciteit, observatievermogen, redeneerkracht, vermogen tot introspectie, of neigingen tot geloofsherziening indien geconfronteerd met nieuwe informatie.

We bestuderen vervolgens twee soorten actoren in detail, met perfect geheugen en juist zonder enig lange-termijn geheugen, in de context van spelen. We laten zien dat bestaande dynamische logica's kunnen worden 'geparametrizeerd' om diversiteit qua geheugen toe te staan, en geven daarbij nieuwe karakterizeringen van de twee typen actoren, met als resultaat volledige dynamisch-epistemische logica's. Een volgende kenmerkende dimensie van diversiteit van actoren is dat deze, ondanks hun verschillen, vaak succesvol hun gedrag weten te coördineren. Hoewel dit thema prominent aanwezig is in de speltheorie en 'multi-agent systems' in de informatica, heeft het nog nauwelijks aandacht gevonden binnen de logica. We analyseren interactie tussen wezenlijk verschillende actoren in enkele concrete scenario's waarin hun 'types' expliciet logisch worden beschreven.

Tenslotte analyseren we het verschijnsel diversiteit in zijn algemeenheid, en bespreken hoe dynamische logica's moeten worden ontworpen om een volledig beeld te geven van actoren met verschillende vermogens en gewoonten, die hun doelen nastreven, en soms ook bereiken, in essentieel sociale situaties.

内容摘要

本书探讨有关理性主体行为的两个主要问题：偏好的动态性和主体的多样性。

我们主要研究关于偏好的两个方面，即，偏好的“原因”和偏好的“变化”。这两个问题被冯赖特1963年出版的重要论著《偏好逻辑》和他的众多后继者们长期搁置，无人问津。本书讨论和比较了偏好的各种概念，并揭示了这些概念之间的联系。特别是，我们集中研究了所谓的外在偏好。与内在偏好不同，外在偏好是基于某种原因的，即，一个人偏好此物而非彼物是有原因的。我们给出了逻辑模型，讨论了这些模型的性质，并提出了关于动态偏好变化的一些逻辑系统和形式结果。

偏好产生于对事物的比较。我们可以对不同的情形进行比较。抽象而言，在这种情况下偏好是介于命题之间的，这里的命题被看作是可能世界的集合。原因则基于可能世界上的“更佳”关系。命题偏好是对这一初始关系的一个提升。我们采取经典的模态逻辑方法，在语言中使用模态词表示更佳关系。然后，我们使用动态认知逻辑的技巧研究偏好的变化，即，一个典型的行动（譬如一个建议或一个命令）可以改变可能世界上更佳关系的次序，从而改变命题偏好。我们给出了动态规约公理来准确地刻画这种变化是如何发生的，从而得到一个完全的动态偏好逻辑。

我们也可以对实物进行比较。具体来说，实物的性质常常决定我们的偏好，而这样的性质就成为偏好的原因。受优选论的启发，我们提出了“优先序列”的概念，它是性质的一个排序。我们考察了从优先序列获得偏好的各种方法，尽管我们主要采取的是优选论的方法。我们利用一阶逻辑的片断来描述这样的情况。就动态方面而言，正是优先序列的改变导致偏好的变化。再次利用动态认知逻辑的方法，我们获得了一组完全的规约公理，可以用来刻画优先序列上可能的动态运算。

无庸置疑，上面的两种观点是紧密联系的。毕竟可能世界本身可以看作是实物！基于对这两种观点的系统比较，我们发展了所谓的“双层视角”，其中模型本身是有层次结构的。特别是，我们证明了介于可能世界层次上的变化和优先序列层次上的变化之间的一些对应结果。最后，我们简要介绍了一个双层偏好谓词逻辑。这个逻辑可以描述更为复杂的情境，在那里我们需要对情形和实物同时进行比较。

但是，故事远没有结束。偏好并非孤立存在，它常常跟知识、信念等认知概念混合在一起。关于混合的方式我们可以有不同的直观，本书讨论了一些可能的选择，并给出了相应的逻辑模型。若在这样的背景下考虑动态性，我们看到的是知识更新、信念修正和偏好改变共同发生，就像在实际生活中那样经常在不知不觉中进行。

以上讨论的最后图景是一个主体以不同的方式处理信息，调整她的信念和偏好。对此，在逻辑上没有指定的唯一规范。本书的第二部分以主体的多样性这个一般现象作为研究重点，因为它给逻辑系统以及这些系统通常预设的理想主体提出了很多挑战性问题。在实际生活中，主体的认知能力和认知习惯在很多方面可以有所不同，例如：她们的记忆力，观察能力，推理能力，内省能力以及面临信息时的修正策略。

我们在博弈的情境中系统地研究了两类主体，即，完美记忆主体和无记忆主体。我们揭示了如何对目前的动态逻辑“参数化”以便容许主体记忆的多样性，从而在动态认知逻辑中获得了对主体类型的一个新刻画。主体多样性在现实生活的另一个方面的体现是，不管主体之间有多大的区别，她们常常能够成功地彼此合作。这是博弈论和多主体系统研究的一个突出主题，但是却很少受到逻辑研究领域的关注。通过分析一些明确反映主体类型的具体实例，我们探讨了不同主体之间的互动。

最后，我们对主体多样性问题做了更为一般性的分析。讨论了什么样的动态逻辑能够对具有不同能力和不同趋向的主体做全面的阐述。这样的主体常常在不可规约的社会环境中探寻她们的目标，并且有时候能够实现她们的目标。

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