Synthese Library 354

Fenrong Liu

Reasoning about Preference Dynamics

Springer

Reasoning about Preference Dynamics

SYNTHESE LIBRARY

STUDIES IN EPISTEMOLOGY, LOGIC, METHODOLOGY, AND PHILOSOPHY OF SCIENCE

VOLUME 354

For further volumes:
http://www.springer.com/series/6607

Reasoning about Preference Dynamics

by

Fenrong Liu
Tsinghua University, Beijing, China

Springer

Assoc. Prof. Fenrong Liu
Tsinghua University
Department of Philosophy
100084 Beijing
Haidian District
China PRC
fenrong@tsinghua.edu.cn

ISBN 978-94-007-1343-7 e-ISBN 978-94-007-1344-4
DOI 10.1007/978-94-007-1344-4
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2011928291

For Huanhuan

Preface

Studies of rational agency have become a major theme in many disciplines, making logic meet with philosophy, computer science, game theory, cognitive science, and other fields. So far, logical theories have mostly described the information that agents have about relevant situations, how it flows, and brings about knowledge update and belief revision. But typically, human beings act on their goals. To fully understand their behavior, we need to take both information and evaluation, the two main driving forces of rational agency, into account. The aim of this book is to provide a clearer picture of how these two crucial forces show analogous static and dynamic logical structure, and also, how they can live in harmony, entangled in many ways.

To achieve this aim, the present book proposes a uniform logical theory of preference, drawing together ideas from several areas: modal logics of betterness relations, dynamic epistemic logics of information change, and priority-based systems for representing structured preference relations. We develop a *two-level view* of preference that fits well with realistic architectures of agency, closer to cognitive reality. But perhaps the key idea underlying this book is *dynamics*, the systematic logical study of acts and events that change information, or agents' evaluation of the world, changing their preferences eventually. The result of our study is a formal framework that has interesting theoretical features of its own, but also, at least as importantly, has the potential of being applied in analyzing preferences in a wide variety of fields. The book provides some first case studies of deontic reasoning and of games to show its theory at work.

Overall, in this book I hope to provide readers with a broad view of the structure of preference, and I argue that changes in preference should be taken on board as an essential part of the enterprise of understanding agency in its proper generality.

Acknowledgements

I thank Johan van Benthem for his constant support throughout this project. I thank Vincent Hendricks for his encouragement in publishing the book in the Synthese Library Series.

I thank the following persons for providing inspiration and useful comments at various stages of this research: Thomas Ågotnes, Natasha Alechina, Guillaume Aucher, Alexandru Baltag, Johan van Benthem, Jan Broersen, Cédric Dégremont, Hans van Ditmarsch, Ulle Endriss, Peter van Emde Boas, Patrick Girard, Sujata Ghosh, Davide Grossi, Till Grüne-Yanoff, Meiyun Guo, Jiahong Guo, Dick de Jongh, Fengkui Ju, Sven Ove Hansson, Andreas Herzig, Tomohiro Hoshi, Zhisheng Huang, Jérôme Lang, Brian Logan, Emiliano Lorini, John-Jules Meyer, Eric Pacuit, Rohit Parikh, Gabriella Pigozzi, Floris Roelofsen, Olivier Roy, Jeremy Seligman, Sonja Smets, Leon van der Torre, Fernando Velazquez-Quesada, Frank Veltman, Yanjing Wang, Tomoyuki Yamada, Zhiguang Zhao, and Jonathan Zvesper. Many chapters of this book have been presented in conferences held in China and internationally, I thank all audiences for their questions and comments. In particular, I thank Yanjing Wang, Junhua Yu, Zhiguang Zhao, and the anonymous referee for their careful reading of the manuscript and for their valuable comments and corrections, which have improved the book a lot.

I thank my Tsinghua colleague Daniel Bell for his kindness and sharing his office with me in the past three years.

I thank Ingrid van Laarhoven and Ties Nijssen at Springer in Dordrecht for their kind assistance: It has been a great pleasure working with them.

Finally, I acknowledge support from the Project (Grant No. 09YJC7204001) and the Scientific Research Foundation for Returned Overseas Chinese Scholars, State Education Ministry in China.

Beijing, China
January 2011

Fenrong Liu

Contents

Part III Preference over Worlds

Part I
Introduction

Chapter 1
Introduction

Humans are often said to be information-processing agents navigating a complex world with their knowledge and beliefs. But *preference* is what colors our view of that world, and what drives the actions that we take in it. Moreover, we influence each other's preferences all the time by making evaluative statements, uttering requests or commands, in ways that direct our search for information, and for actions that best fit our goals.

A phenomenon of this wide importance has naturally been studied in many disciplines, especially in philosophy and the social sciences. This book takes a formal point of view, being devoted to logical systems that describe preferences, changes in preference and entanglement of preference and belief. I will plunge right in, and immediately draw your attention to the first time when preference was extensively discussed by a logician.

1.1 A Brief Historical Sketch of Preference Logic

1.1.1 Von Wright and the Basic Ideas

In his seminal book *The Logic of Preference: An Essay* from 1963, Georg-Henrik von Wright started a full-scale logical study of notions that interest moral philosophers. He charted this space with the following three dimensions, which of course admit border-line cases:

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

The intuitive concept of preference was said to stand between the last two of these groups: 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. And as we shall see later in this book, in a logical perspective, preference also fits with deontology.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_1,

Von Wright chose to make a primitive notion of preference rather than betterness the starting-point of his inquiry.[1] In particular, he went on to define a logical notion of preference as a *relation $P\varphi\psi$ between propositions* which states intuitively that the agent finds every instance of φ better than every instance of ψ. Von Wright then observed how the resulting notion satisfies a number of formal laws of reasoning, such as transitivity ("$P\varphi\psi$ and $P\psi\chi$ imply $P\varphi\chi$") or monotonicity ("$P\varphi\psi$ implies $P(\varphi \wedge \chi)\psi$"). Starting from this foothold, reasoning with preference was drawn into the scope of logic. A broad stream of studies on preference logic ensued, which shows no sign of diminishing. We will not use von Wright's system in this book, but some modern modal versions. For details of the original system, we refer to the Handbook chapter [101].[2]

We will soon say more about later developments in preference logic. But for the purposes of this book, we immediately note two further themes in von Wright's work that stand out, precisely because he did not develop them, and they did not become mainstream topics. Still, they seem crucial to the functioning of preference.

Extrinsic versus intrinsic preference First, while considering the relationship between preference and betterness of worlds or objects, von Wright distinguished two kinds of relation: *extrinsic* and *intrinsic* preference. 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. . . .
>
> ([197], 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.

The division between intrinsic and extrinsic is by no means the only natural way of distinguishing things. One can also study varieties of moral preference, aesthetic preference, economic preference, etc. But reason-based preference seems as fundamental and "logical" as intrinsic preference, even though von Wright did not return to it in his later systems.

A first main goal of the present book is to fill this gap, by extending the literature on intrinsic preferences with formal logical systems for the *extrinsic notion of preference*, allowing us to spell out explicit reasons. On the way there, we will also make new contributions to the literature on intrinsic preferences, since we see this as a case of extension, not replacement.

---

[1] The earlier study [93] did propose logic systems for the notion of betterness.

[2] Incidentally, von Wright's reading for the above $P\varphi\psi$ has a rider "other things being equal", meaning that one must keep the truth values of certain relevant predicates constant. We will discuss these *ceteris paribus* aspects of preference later on.

Preference change A second major phenomenon observed by von Wright, and then immediately removed from his logical agenda, is found in the following quote:

> 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.
>
> ([197], p. 23)

But clearly, our preferences are not static! One may *revise* one's preferences for many legitimate reasons, and indeed, rational agency consists partly in dealing with this flux. And just as with information, it is then the interplay of reasoning unpacking current states and acts that generate new states which should be explained.

The second main issue dealt with in this book is therefore how to model preference change in formal logics. This leads to new dynamic versions of existing preference logic, and interesting connections with other areas showing a similar dynamics, such as belief revision theory.

Following von Wright's work, many studies on preference logic appeared over the last few decades, and those not just in philosophy. Due to its central character, at the interface between evaluation, choice, action, moral reasoning, and games, and even computation by intelligent agents, preference has become a core research theme in many fields. In what follows, I will summarize some main issues, just to put this book in perspective. My purpose is not to give an overview of the vast literature (I give some basic references for that), but only to show how this book is grounded in a historical tradition. I point out some issues from the literature that are relevant to the present book, and I end by highlighting some particular proposals that have inspired the main new themes developed in it.

1.1.2 Preference in Logic and Philosophy

Formal investigations on preference logic have been intensive in philosophical logic. The best survey up to 2001 can be found in the Chapter *Preference Logic* by Sven Ove Hansson in the *Handbook of Philosophical Logic* [101].

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 [198]. 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.
>
> ([101], pp. 346–347)

Most philosophical logicians have concentrated on exclusionary preferences. However, in this book we will consider both. As we will show later, 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 relations.

Another much-debated issue has been whether certain *principles* or "structural properties" are reasonable for preference. Here economists joined logicians, to discuss the axioms of rational preference. Many interesting scenarios have been proposed that argue for or against certain formal principles, resulting in different logical systems disagreeing in their basic principles of reasoning (cf. [124, 134, 167, 187], etc.). However, a general critical result in [95] is worth noticing. 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 after all. We will sidestep these debates. In general, our logical systems do 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. [76, 112]).

There are also obvious relations between preference and *moral* or more generally, *evaluative* notions like "good" and "bad". The issue then arises which notion is the more primitive one. Several researchers have suggested definitions for "good" and "bad" in terms of the dyadic predicate "better".[3] In this line, [97] 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. Reference [22] defines binary comparatives like "better" in terms of context-dependent predicates "good", and [160] takes this much further into a general analysis of comparative relations as based on a "satisficing" view of achieving outcomes of actions.[4] This book will follow the line that takes betterness as primitive, although one might say that our later analysis of extrinsic preference as based on unary constraints has some echoes of the linguistic strategy deriving binary comparatives from unary properties.

Finally, again within philosophy, there has always been a strong connection between preference and moral reasoning. This is clear in *deontic logic*, another branch of philosophical logic going back to von Wright's work, this time to his [196]. 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 have been introduced in deontic logic to interpret both absolute and conditional obligations (cf. [98, 183]). Preference also helped solve some persistent "deontic paradoxes". Reference [59] gave a deontic interpretation of the calculus of intrinsic preference to solve the *problem of supererogation*, i.e., acting beyond the call of duty. In another direction, [184] extended the existing temporal analysis of *Chisholm Paradox* of conditional obligation in [71] using a deontic logic that combines temporal and preferential notions.

---

[3] A widespread idea is to define "good" as "better than its negation" and "bad" as "worse that its negation", as in [197] and [93]. Quantitative versions of this are found in [125].

[4] It would be of interest to contrast their "context-crossing principles" with Hansson's views.

Finally, we mention [185], which improved solutions to deontic paradoxes by combining preference logic with updates from the "dynamic semantics" of natural language.

We too will look at deontic applications of our systems later on in this book, since moral reasoning is a powerful area of concrete intuitions for preference representation and preference dynamics.

1.1.3 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 orderings [95]. Moving back to logic, [159] 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 aspects of these concepts, i.e., the possibility of measuring them.[5] In terms of axiomatizations, the standard approach takes weak preference ("better or equal in value to") as a primitive relation (cf. [170]), as we will also do in this book.

Of course, economists have also added further themes beyond what had been considered by philosophers. One prominent case are tight connections between *preference* and *choice* [170, 171]. Preference is seen as "hypothetical choice", and choice as *revealed preference*. Recently, revealed preference has become prominent in understanding the concept of equilibrium in game theory (cf. [111]). Preference is then no longer a primitive, but a notion constructed out of observed outcomes. The observed outcomes of behaviour can be rationalized by postulating preferences realizing one's chosen equilibrium notion for decisions or games.

Of course, views of preference need not agree across fields. For instance, already [197] pointed out that "it is obvious that there can exist intrinsic preferences, even when there is no question of *actually* choosing between things" ([197], p. 15). In this book, we will not take sides in this dispute, though we will discuss the issue of revealed preference in our Chapter 12 that takes a look at game theory from the perspective of our preference logics.[6]

1.1.4 Preference in Computer Science and Artificial Intelligence

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

[5] Modern studies in this line are [52] and [186].

[6] Another congenial area of formal studies on preferences for agents, and in particular, how these can be rationally merged, is *Social Choice Theory*: cf. [75].

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, [60] studied general principles that govern agents' reasoning in terms of their belief, goals and actions and intentions. This inspired the well-known "BDI model" in [155], which shows how different types of rational agents can be modeled by imposing conditions on the persistence of agents' beliefs, desires or intentions.[7]

Interestingly, notions from von Wright's work have made their way directly into the computational literature on agency. In particular, his idea that preferences can often only be stated *ceteris paribus* has been taken up in [66] and [68], which studied preference "all else being *equal*". The other main sense of ceteris paribus, as "all else being *normal*", was taken up in [54], 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 [39]. The eventual systems of our book can deal with the normality variant, as we will show in our Chapter 5 on doxastically entangled preference.[8]

Preference occurs in many other strands in the computational literature, but we will not list of all of these here.[9] In the course of our chapters, it will become clear how methods from the computational logic tradition have influenced this book.

1.2 The Main New Themes in this Book

Our starting point was the preference logic of [197], its general methodology, and some major conceptual distinctions that von Wright made. We also saw two major issues that he noted, but left out of his logical analysis, viz. *reason-based extrinsic preference*, and the *dynamics of preference change*. These will be the two main new themes in this book. But on the side, we will also have a third theme, again not explicitly taken up by von Wright, viz. the *entanglement of preference and belief*, illustrating the crucial interplay of information and evaluation dynamics in successful rational agency.

The main point of this book is to show how these crucial aspects of reasoning with preference can be treated in a uniform logical framework, which brings together ideas from several different areas: (a) the subsequent development of preference logic, (b) the computational literature on agents, and (c) recent developments

[7] Modern developments are found in [107, 129], and [195].

[8] Adding the equality variant seems feasible, too, but we will not pursue it.

[9] Some sources are the AI literature on circumscription (cf. [173]), qualitative decision theory for agents planning actions to achieve specified goals (cf. [55, 67, 182]), and recent computational studies of efficient preference representation (cf. [57, 62] and [188]). Preferences also occur in the core theory of computation, e.g., in describing evolutions of systems using some measure of "goodness" of execution: cf. [141] and [172].

in the theory of belief revision and dynamic epistemic logic. In what follows I briefly introduce my guiding intuitions, and the main ideas.

1.2.1 Richer Representations: Reasons for Preference

In many situations, it is quite natural to ask for a reason for stated preferences. As von Wright says in his discussion of "extrinsic preferences",

> 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 one's health is the reason for the preference for claret over hock. Similar examples abound in real life: Say, I prefer one house over another *because* the first is cheaper and of better quality than the second. As the second example shows, such reasons may be of many kinds, since the criteria that determine preference can be diverse. In line with this, reasons can be of various kinds: from general principles to individual facts about objects. More generally, preference is often a structured multi-criterion notion, as one can see in many areas, from economics (cf. [161]) to linguistics (cf. [154] on optimality theory, where grammatical analysis means choosing most preferred sentence readings).

All this also suggests a task for *logic*. Giving reasons is an eminently logical task, making things susceptible to reasoning. We will therefore give logical models for reason-based preference that have both a "betterness" order among worlds or objects, and a structure of reasons inducing that betterness order, in the form of so-called "priority graphs". We will show how this richer format of representation fits well with the existing tradition in preference logic, but also adds greater depth of analysis for many old and new issues.

But giving reasons and reasoning with them at once brings in further issues:

1.2.2 Entanglement: Preference, Knowledge and Belief

Reasoning about preferences often involves agents' *information*. I may prefer a certain object to another right now, because I do not yet know about some decisive flaw. I may prefer taking an umbrella (despite the inconvenience of carrying it), if I believe that it is going to rain. And in a similar vein, people's obvious diversity qua preferences is matched by their obvious diversity qua beliefs.

This "entanglement" of preference with information-based attitudes like belief, or knowledge, seems essential to agency. This is of course not a new insight. Entanglement is the norm in decision theory and game theory when modeling decision making under uncertainty ([113, 166]). And even more generally, a crucial notion underlying probability theory is *expected value*, whose definition mixes probability and numerical utility, the quantitative correlates of belief and preference.

Now entanglement also poses an obvious logical challenge. Informational attitudes like knowledge and belief have been extensively studied: How do these

systems interface with preference logic? We will show how this can be done, finding various "degrees of entanglement" for informational and evaluational attitudes, in a way that relates recent developments in various fields.

Finally, the preceding two themes strongly suggest a third, that we will introduce now. It might be the main innovation in this book.

1.2.3 Preference Change

It is easy to observe that asking for reasons is often a prelude to argumentation, where one tries to *change* someone else's views. Indeed, merely giving information can change preferences, witness the following twist to our von Wright's scenario:

> Suppose that before seeing his 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!

This seems so important and natural that we want a logical account of this preference change. But we will also go one step further. Recall the earlier distinction between extrinsic and intrinsic preference. The latter can change too. I can fall out of love with one of my two suitors without having to give a reason, my evaluation of that person has just changed. Thus, preference change between objects does not have to come about because some underlying structure changed: It can just be a direct change in my tastes or feelings, as reflected in the way I order my alternatives. This kind of change, too, should be described.

I will present a logical theory of both extrinsic and intrinsic preference changes. My solution is in line with a "dynamic turn" that has already occurred in recent studies of knowledge update and belief revision.[10] In particular, we can use current models of information change to also deal with preference change, drawing on analogies with recent dynamic logics of information flow (cf. [14, 18, 29]).

1.2.4 The Total Theory

The logical theory developed in this book proposes a uniform framework for representing rich forms of intrinsic and extrinsic preference, as well as their entanglement with knowledge and belief. Moreover, it treats statics and dynamics on a par: Acts of preference change are an explicit part of the logic, allowing us to track an agent's informational and evaluational behavior over time.

With this theory in place, it is natural to return to some of the traditions lightly reviewed in the above. One obvious related area is belief revision theory where our techniques make a lot of sense.[11] In this book, I will also discuss repercussions in

---

[10] In this sense, the above entanglement of preference and belief also has a useful aspect, as a source of analogies.

[11] Reference [18] on "safe belief" and its dynamic axioms acknowledges [42], a pre-study for this book.

the areas of deontic logic (cf. Chapter 11 on changes in obligations and in norms) and game theory (cf. Chapter 12 on entanglement in solution procedures, as well as preference dynamics in games), where our ideas return in concrete settings leading to many new questions.

1.3 Guide for the Reader

This book is aimed at people from various disciplines who are interested in studying agency with logical models. We assume some prior acquaintance with epistemic and dynamic logic (though we will summarize some basics), and we write for readers familiar with the basics of first-order logic and modal logic. Here is a more detailed description of the parts that are to follow:

1.3.1 Part II: Dynamics of Information

Chapter 2. Dynamic Epistemic Logic This chapter is a review of epistemic logic, and especially, dynamic epistemic logic, to set up the methodology that we will later use for preference in this whole book. As for dynamics, we mainly look at "public announcement logic", with a focus on the technical details that will play a role in later chapters when we consider dynamics for preference. The more powerful method of "product update" is introduced lightly, too. Overall, we stress features of dynamic epistemic logic that are more broadly important from a methodological point of view.

1.3.2 Part III: Preference Over Worlds

Chapter 3. Preference Over Worlds: Static Logic A modal semantics is given for static intrinsic preferences. Models consist of a universe of possible worlds, representing the different relevant situations, endowed with a primitive binary order of "betterness".[12] This ordering may be connected (any two worlds stand in some betterness relationship), but in general, we allow incomparable cases, and work with *pre-orders* that are merely reflexive and transitive. Pre-orders set the level of generality aimed for throughout the later chapters of this book. We give complete axiomatizations from the literature for the modal logic of extrinsic preferences. In addition, we discuss other basic themes, such as "generic preferences" between propositions. This requires a study of "lifting" betterness order from worlds to sets of worlds, with stipulations such as the $\forall\exists$-rule, which says that every φ world has at least one better ψ alternative world. These lifts, and many other types of statement can be described

[12] Note how this makes betterness a preference over incompatible alternatives.

in our standard modal language over betterness models. As an illustration, we study the ∀∃-lift, and explore its logical properties in more detail.

Chapter 4. Preference Over Worlds: Dynamic Logic This chapter provides a format for studying preference dynamics in the appropriate generality. We start from a simple test scenario that may be called a "suggestion". Statements like suggestions or commands *upgrade* agents' preferences by changing the current betterness order among worlds. A dynamic betterness logic is presented that represents such actions explicitly, and then axiomatizes their complete theory using dynamic-epistemic-style reduction axioms. We show how this system automatically describes changes in generic preferences. Beyond specific examples, we present a general format of "relation transformers" for which dynamic epistemic reduction axioms can be derived automatically. As an illustration, we show how this system can handle default reasoning and various policies for belief revision.

Chapter 5. Entanglement of Preference, Knowledge and Belief Preference, knowledge and belief occur intertwined. This chapter investigates possible ways of entangling these notions. We first explore a simple manner of combining knowledge and preference languages, viz. juxtaposition of epistemic logic and preference logic. We look at how this simple combination affects the dynamics, and we propose a new update mechanism: update by "link-cutting". Next, we show how this method can be applied to the combination of preference and beliefs. Then we study a more intimate way of combing preference and beliefs, with new merged "intersection modalities", proposing models and a complete logic. In the resulting language, we can talk of preference and plausibility in much richer combinations. Finally, we show that the new entangled models lead to richer notions of generic preference, too.

Chapter 6. Intermezzo: A Quantitative Approach While the main line in this book is qualitative logics, we do make one excursion to numerical models. Our aim here is mainly to show how no insuperable barrier separates the methods in this book from the much larger world of quantitative approaches to preference and action. First, we introduce a complete system of "epistemic evaluation logic" allowing for finer gradations than what we had so far, and we define a matching notion of numerical bisimulation. Then we define a new product update mechanism to deal with the dynamics of preference in this richer framework. We also present the resulting complete dynamic epistemic evaluation logic. As a special topic, we then show how this update mechanism lends itself naturally to "parametrization": allowing us to model the phenomenon of *diversity of agents* which is so characteristic of real life. We end with some comments on deontic reasoning in this setting.

1.3.3 Part IV: Preference from Priorities

Chapter 7. Preference from Priorities: Static Logic Now we move to richer models for preference structure, as required by extrinsic preferences. This time, the primary scenario is preferences over objects, again standing for incompatible alternatives.

For this purpose, inspired by linguistic optimality theory, we introduce linear "priority sequences" of *constraints*, i.e., relevant properties of objects. Through the natural induced lexicographic ordering, priority sequences supply reasons for preference by comparing objects as to the properties they have or lack in this sequence. We also determine the power of this method through structural representation theorems.[13] Intuitively, these theorems say that one can always find a reason for some given betterness pre-order. A complete preference logic is proposed and a proof of a representation theorem for the simple language is presented. Finally, we make an important generalization. Linear priority sequences induce connected orders, and while this is an important special case, the proper generality for our preference logic are the earlier *pre-orders*. We show how the general representation for our preference then becomes a framework of "priority graphs", inspired by the seminal paper [6], which provides a natural and eventually more realistic generalization of our earlier account.

Chapter 8. Belief-Based Preference In the real world, agents only have incomplete information. Therefore, we now add epistemic and doxastic structure to our structured preference models. In particular, we introduce beliefs that help form preferences. We propose three different ways of defining preference in terms of priorities and beliefs. In particular, we present a doxastic preference logic for the notion of "decisive preference" and prove a representation theorem for that case. Next, we extend our discussion to the multi-agent case, where we study both cooperative and competitive agents, and again capture their characteristics in the form of representation theorems. Moreover, we look at generic preference over propositions in this context, and propose a propositional doxastic preference logic. Finally, we discuss connections between preference over objects and preference over propositions.

Chapter 9. Preference from Priorities: Dynamic Logic This chapter explores dynamics in the present richer static setting. Changes of preference can now have two different reasons: either changes in priority sequences or priority graphs, or changes in belief structure inducing preference changes through entanglement. For the latter, we describe two cases: Belief changes due to "hard information", and belief changes due to "soft information". Again, the general dynamic-epistemic approach works here, treating the relevant events triggering change as dynamic actions that are subject to recursive reduction axioms.

1.3.4 Part V: A Two-Level Perspective on Preference

Chapter 10. A Two-Level Perspective on Preference In this chapter, we draw our two earlier approaches together: intrinsic modal betterness dynamics and priority dynamics. We propose a "two-level model" for preference structure and its dynamics. This merges ideas from earlier chapters, but with some new twists. For instance,

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

we need to extend the general "program format" of Chapter 3 to the strict betterness relations used in the priority graphs of Chapter 7. We also lift the representation theorem for priority sequences to the latter more general setting, representing the more realistic case of possibly incomparable or even conflicting priorities. We then turn to connections between the two dynamics, at the level of worlds and at the level of propositional priority structure. We prove a general inclusion result from "basic graph priority changes" to *PDL*-definable modal transformers on betterness models. But we also find examples that show that there is no total inter-level reduction of perspectives, either way. Finally, we discuss what this means for the interplay of extrinsic and intrinsic preferences. In particular, we find that intrinsic preference can become extrinsic when we are willing to allow for a new form of dynamics, less studied in dynamic logics so far, viz. *language change*. In all, having both levels around in preference logic is important for two reasons. It allows for more realistic and sophisticated modeling of preference scenarios, but it is also a source of interesting new notions and open problems.

1.3.5 Part VI: Applications and Discussions

Chapter 11. Deontic Reasoning In this part, we confront our ideas with some major areas where preference plays an important role. This first chapter shows how the static and dynamic preference logics developed in the preceding parts can be applied concretely to deontics. One key illustration are "contrary-to-duty obligations", interpreted in terms of our reason-based priority graphs. We then discuss the perspective on deontics obtained by juxtaposing the semantic view of standard deontic logic in terms of betterness relations with the syntactic view of structured references. We also apply our other main theme of preference change. In particular, we use the techniques of Chapter 10 to study deontic betterness dynamics, through a correspondence between "syntactic" normative changes, and semantic ones at the level of deontic betterness. We show how this throws new light on some old paradoxes of deontic reasoning. Finally, as an appendix, we add some thoughts on the linguistic aspect of normative behavior. Commands are normally uttered in natural language, using "imperatives". A congenial approach to studying the meaning of imperatives is "dynamic semantics" in terms of appropriate state changes for language users. We briefly explore how our ideas can enrich existing dynamic semantics for imperatives.

Chapter 12. Games and Actions In this second applied chapter, we look at game theory, another area where preference is essential to making sense of behavior. Drawing on some recent literature on logic and games, we show how most of main themes play in the concrete compass of solution methods for extensive games such as "Backward Induction". We find an excellent fit for our modal preference logics over betterness models, entanglement of preference with information in standard notions of *rationality* for players, and especially, information dynamics for both knowledge and belief in the actual process of game solution. We link these with various logical issues that have played in our framework: For instance, diversity of

players turns out to connect with our interest in options for lifting betterness to set preferences. Toward the end of the chapter, we discuss what the new topics of this book might do, when thrown into this lively current research area. In particular, we look at uses of *preference change* in extensive games, focusing on procedures for *rationalizing* given behavior. We also briefly explore endowing games with richer *priority structure* for players' goals, and its possible consequences for notions of strategic equilibrium.

1.3.6 Part VII: Finale

Chapter 13. Conclusion We conclude with a summary of the logic of preference as developed in this book. After that, we briefly discuss where we see the main lines ahead, namely, a treatment of group agents, long-term temporal processes, links with probability and other quantitative features, and finally, more fine-grained syntactic representations for preference.

1.4 Some Major Influences on This Book

The ideas in this book stand in a long tradition. In addition to the brief history that we sketched earlier, we state a few particular links to earlier work.

First, the *modal preference logic* in this book started at least with [55], and then [94]. The dissertation [83] is a modern version congenial to ours. These logics set the pattern for our static base systems, though we have also incorporated some ideas from the recent study [39].

Second, as to *reasons for preference*, even though most authors have concentrated on intrinsic preferences, there are exceptions, witness the brief survey in [103]. Our initial representation with priority sequences owes a lot to linguistic optimality theory [154] which describes successful language use in a rule-free manner, in terms of optimal satisfaction of syntactic, semantic, and pragmatic *constraints*. Our eventual priority graphs are a specialization of a general mathematical framework for belief merge and social choice proposed in the elegant algebraic study [6].

Third, the *entanglement of preference and belief*, too, has a long history. For instance, [123] proposes a logic of "desires" whose semantics contains two ordering relations of preference and normality, with desires referring to the best of one's most normal worlds. This is a typical example of the sort of entanglement that we have analyzed. Another source are modal preference logics of games, that we acknowledge separately below.

Four, our treatment of *preference change* has many ancestors, closer or more distant. First, the idea that preference change might be formulated in dynamic logic occurs as early as [33].[14] More in the style of AGM belief revision theory (cf. [2]),

[14] The idea that *propositional dynamic logic of programs* (cf. [104]) is relevant to deontics can even be traced back to [142].

[100] proposed postulates for four basic operations in preference change. This book, however, has taken the "model construction" approach of *dynamic epistemic logic* (*DEL*), something which will be obvious to any reader who is familiar with that paradigm. There the emphasis is on logics for concrete dynamic actions that change agents' information and their corresponding attitudes. Some key sources here are [14, 65], and [32].[15] The idea that preference change can be dealt with in terms of these methods also occurs in recent work by Tomoyuki Yamada, of which [200] is a representative sample.

Finally, we mentioned some related sources that have been relevant to how we see our links to neighboring areas. We have mentioned several papers linking preference and deontic logic already, such as [98, 183]. But other contacts with philosophy are relevant, too, and we see our work as similar in spirit to logical studies in the philosophy of action, such as the dynamic logic of action in [203], or the logics of action and intention in games of [164].[16] Moving beyond philosophy proper, social choice theory is a rich source of logics for preference aggregation by collective agents, and [189] is a good sample of ideas that form a natural continuation of ours. Next, [161] is a pioneering study on broader contacts between logic, belief revision and economics. Finally, as for our special focus on game theory, some of our main leads have come from [46, 82].

Published sources for this book The material in this book comes from the dissertation [131] plus a number of follow-up publications. Chapters 3 and 4 go back to [42]. Chapter 6 elaborates some ideas from the master's thesis [130], and [133]. Chapters 7, 8, and 9 go back to [117]. Chapter 10 is based on the paper [135]. Chapter 11 is a version of [48], presented at DEON 2010. The appendix on imperatives is the working paper [119]. Chapter 12 is new to this book, though I must thank the authors of [46] for letting me use their unpublished paper as a red thread in the first half. I thank my various co-authors in these publications for their kind permission to use the material here.

[15] Some further publications for what is sometimes called the "Amsterdam approach" are [23, 81, 152, 192], as well as the work on belief revision in [10, 29] and [18].

[16] The philosophy of action contains many further instances of ideas in this book. E.g., its distinction between "recognitional" and "constructivist" views of practical reasoning [193] mirrors our distinction between intrinsic and intrinsic preference.

Part II
Dynamics of Information

Chapter 2
Dynamic Epistemic Logic

2.1 Introduction

What is knowledge? How can we acquire knowledge? When can we say that we know something? Do we know that we know something? Those are the issues that puzzled Chinese philosophers about 2000 years ago, witness a famous dialogue below between Zhuangzi (approx.369–286 BC) from the Daoism School and Huizi (390–317 BC) from the School of Names:

> *One day Zhuangzi and Huizi are strolling on Bridge Hao.*
>
> *Zhuangzi*: "*Look how happy the fish are just swimming around in the river.*"
>
> *Huizi*: "*How do you know they are happy? You are not a fish.*"
>
> *Zhuangzi*: "*And you are not me. How do you know I don't know the fish are happy*?"
>
> *Huizi*: "*Of course I'm not you, and I don't know what you think; But I do know that you're not a fish, and so you couldn't possibly know the fish are happy.*"
>
> *Zhuangzi*: "*Look, when you asked me how I knew the fish were happy, you already knew that I knew the fish were happy. I knew it from my feelings standing on this bridge.*"

This is a typical interaction between two agents (or more, if you count the fish in), in our modern jargon. The core issue under discussion is the following: When can we say that one agent knows another agent's feeling or his knowing something?

Similarly, in the West, the nature of knowledge was discussed already in Plato's *Theaetetus* and works by Aristotle, and the debate concerning the definition and scope of knowledge has been going on ever since. The research field concerning the above knowledge-related questions is generally called "epistemology". Many fascinating issues have been studied, for instance, the structure of knowledge itself (cf. [58, 151] and [180]), the relations between knowledge, belief, justification, and evidence (see e.g., [61] and [85]), or the debate between internalism and externalism [69, 146].[1]

There is a vast body of philosophical literature here that is worth our attention. But for the purposes of this book, we pick out only a few of these strands, and turn to a more formal or logical perspective.

[1] Reference [11] studied epistemic logic from both internal and external perspectives.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_2, © Springer Science+Business Media B.V. 2011

In the circle of logic, Hintikka's pioneering work [115] (*Knowledge and Belief: An Introduction to the Logic of the Two Notions*) from 1962 analyzed the notions of knowledge and belief for individual agents by means of modern formal logics for the first time. He based his account on the possible world semantics which was getting popular at that time. Later on the notion of common knowledge for groups of agents was added in Lewis' work [126]. The study of epistemic logic was under way then, and various attempts have been made to capture the philosophical intuitions that we have about knowledge. The notions of knowledge and belief also gained attention in computer science and AI, as these subjects started the study of computation in groups of intelligent agents: See [143] and [73]. The result was a far richer agenda of topics and techniques for epistemic logic. And one more discipline entered this mix in the mid-seventies. Knowledge and belief were studied in game theory too, as game play involves players' individual and common knowledge and beliefs. A classic paper starting this interface of logic and game theory is [156].

Over the last three decades, information has come to play a big role in many different fields, and philosophy and logic are no exception. And in line with this, the information-driven dynamics of knowledge and belief has come to the fore. Various proposals have been made in this regard. For instance, the *AGM*-paradigm was put forward in the 1980s [2] and it described how to revise one's beliefs when new information comes in while remaining consistent. Non-monotonic logics (cf. [140, 157], etc.) started around the same time, with one of its goals being to incorporate new incoming information that might be an exception to some initial rule. While these approaches are well-known by now, in this chapter we are going to introduce one more recent paradigm, that of *dynamic epistemic logic* (*DEL*). This has slowly emerged since the 1990s (cf. e.g., [14, 32, 152], and [81]), and it aims to provide a general logical mechanism for dealing with new information.[2] In doing so, *DEL* starts from basic epistemic logic, now reinterpreted as a theory of what has been called "semantic information". By now, *DEL* has become a very powerful engine to handle knowledge and belief changes. In this book, this approach will be adapted for the case of evaluation dynamics, resulting in changes in agents' preferences.

To set up the basic background for our further investigation, we review the basics of dynamic epistemic logic here. Please note that our treatment will be very brief and mostly methodological. For complete didactic details and a genuine textbook-style introduction, we refer to the cited sources.

2.2 Epistemic Logic

In what follows we immediately introduce the standard definitions of epistemic models and logic, adopted from the previous literature (e.g., [73] and [50]).

[2] References [65] and [32] are two comprehensive and more up-to-date references.

Definition 2.1 (epistemic language) Let a set of propositional variables Φ, a finite set of agents N be given. The epistemic language is defined by

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \quad \text{where } p \in \Phi, a \in N.$$

The language of epistemic logic is an extension of that of propositional language. We follow the usual abbreviations in propositional logic:

$$\varphi \vee \psi := \neg(\neg\varphi \wedge \neg\psi).$$

$$\varphi \rightarrow \psi := \neg(\varphi \wedge \neg\psi).$$

$$\varphi \leftrightarrow \psi := (\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi).$$

For the knowledge operator $K_a\varphi$, we write $\langle K_a \rangle$ to be its dual and the relation between them is the following:

$$\langle K_a \rangle \varphi := \neg K_a \neg\varphi.$$

The intended reading of $K_a\varphi$ is that "agent a knows that φ", and that of $\langle K_a \rangle \varphi$ is that " it is consistent for agent a to know that φ."

Definition 2.2 (epistemic models) An epistemic model is a tuple $\mathfrak{M} = (S, \{\sim_a \mid a \in N\}, V)$, where S is a non-empty set of epistemically possible states,[3] is an equivalence relation $\sim_a$ on S, and V is a valuation function from Φ to subsets of S.

In this context, we interpret the knowledge operator with an equivalence relation and take **S5** as the logic system for knowledge. But this is optional in an approach. There are extensive philosophical discussions about its justification. Various alternatives haven been proposed in terms of model classes. We will not go into the details on this issue here.

Definition 2.3 (truth conditions) Given an epistemic model $\mathfrak{M} = (S, \{\sim_a \mid a \in N\}, V)$, and a state $s \in S$, we define $\mathfrak{M}, s \models \varphi$ (formula φ *is true in* $\mathfrak{M}$ *at* s) by induction on φ:

$\mathfrak{M}, s \models \top$ iff always.

$\mathfrak{M}, s \models p$ iff $s \in V(p)$.

$\mathfrak{M}, s \models \neg\varphi$ iff not $\mathfrak{M}, s \models \varphi$.

$\mathfrak{M}, s \models \varphi \wedge \psi$ iff $\mathfrak{M}, s \models \varphi$ and $\mathfrak{M}, s \models \psi$.

$\mathfrak{M}, s \models K_a\varphi$ iff for all t : if $s \sim_a t$, then $\mathfrak{M}, t \models \varphi$.

[3] In this book, possible states/worlds are denoted by variables w, v, s, t, but also sometimes x, y, as seems convenient in the statement of notions and proofs.

Theorem 2.4 *The epistemic logic EL can be axiomatized completely by the following axiom schemes and inference rules:*

(1) *Tautologies of propositional logic.*
(2) $K_a(\varphi \to \psi) \to (K_a\varphi \to K_a\psi)$.
(3) $K_a\varphi \to \varphi$.
(4) $K_a\varphi \to K_aK_a\varphi$.
(5) $\neg K_a\varphi \to K_a\neg K_a\varphi$.
(6) *If* $\vdash \varphi$, $\vdash \varphi \to \psi$, *then* $\vdash \psi$.
(7) *If* $\vdash \varphi$, *then* $\vdash K_a\varphi$.

The proof of this theorem can be found in any modal logic textbooks (e.g. [50] and [47]). Axiom 2 expresses closure of knowledge under known consequences. This form of logical omniscience can be questioned, but we will stick with it for most of this chapter. Axioms 4 and 5 express what has been called positive and negative introspection.

By now we have presented an epistemic logic characterizing the notion of static knowledge. However, information flows, it keeps changing what we perceive, and we obtain new knowledge over time. Consider the simplest scenario, public announcements, which improve ignorance in social communication.

2.3 Public Announcement Logic

Public announcements abound in real life. News is publicly broadcast through television or radio, making us know what is happening around the world. We often make announcements in ordinary life, like announcing one's marriage to one's friends. As a teacher, we assign homework to the students. Consider the following scenario:

Example 2.5 (homework assignment) After finishing the class, Professor Zhang said: "The homework for today is Exercise 5 on Page 321."

Let p denote the proposition "The homework for today is Exercise 5 on Page 321." Before Professor Zhang said anything, the students did not know whether it is the case that p. Let us consider things from one student a's point of view, her epistemic state can be described in the picture shown in Fig. 2.1:

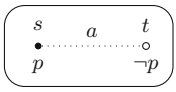


Fig. 2.1 Before a public announcement

In this model, there are two possible worlds, s and t, proposition p holds at s but not at t, s is the actual world. A dotted line between s and t denotes an epistemic uncertainty equivalence relation, i.e. from the perspective of agent a, s and t cannot be distinguished. Reflexive arrows are omitted throughout the text.

However, after the announcement of p, a knows that p, pictured in Fig. 2.2.

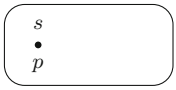


Fig. 2.2 After a public announcement

Now the world t in which p does not hold in the initial model is eliminated, only one world s is left, where p holds. Note that it is a submodel of the original model. Then a knows that p in this new model.

Public announcement logic (*PAL*) is meant to study the logical rules of knowledge change under public announcements as illustrated above. It is a combination of epistemic logic and one kind of dynamic action, namely, *public announcement*. We now define the *language* of *PAL*:

Definition 2.6 (PAL language) Let a set of propositional variables Φ, a finite set of agents N be given. The language of public announcement logic is defined by

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid [!\varphi]\psi \quad \text{where } p \in \Phi \text{ and } a \in N.$$

Compared with the static epistemics language, a dynamic modality $[!\varphi]\psi$ is added, with an intended reading "after announcing φ truthfully, ψ holds."

Following the intuitions we gave for update with announcement, we now formally define the updated model:

Definition 2.7 (updated model) Public announcements of true propositions φ change the current model into its updated model as follows:

Consider any model $\mathfrak{M}$, formula φ is true at some world s. We define the updated model $(\mathfrak{M}|\varphi, s)$ ("$\mathfrak{M}$ *relativized to* φ *at* s)" to be the submodel of $\mathfrak{M}$ whose domain is the set $\{t \in \mathfrak{M} \mid \mathfrak{M}, t \models \varphi\}$.

Definition 2.8 (truth condition) Omitting the standard clauses for the usual operators, the truth condition of the new dynamic formulas is defined as:

$$\mathfrak{M}, s \models [!\varphi]\psi \text{ iff if } \mathfrak{M}, s \models \varphi, \text{ then } \mathfrak{M}|\varphi, s \models \psi.$$

Note that we evaluate the formula ψ at those states where φ holds, this condition is called the *precondition* for $!\varphi$. In the case of public announcement, this simply means that the announced proposition must be true. Typically, informative announcements update models with more uncertainties to a new model with fewer uncertainties. This way of update is often described as "eliminative approach by hard information", as what we have seen in the above scenario. Agents obtain new knowledge after the announcement. This language can make characteristic assertions about knowledge change such as $[!\varphi]K_a\psi$, which states what agent a will know after having received the hard information that φ. The knowledge change before and after the update is characterized in so called *reduction axioms*. Here is the complete logical system of information flow under public announcement (cf. [81, 152]):

Theorem 2.9 *PAL is axiomatized completely by the usual laws of epistemic logic plus the following reduction axioms:*

(1) $[!\varphi]q \leftrightarrow (\varphi \rightarrow q)$ *for atomic facts* q.
(2) $[!\varphi]\neg\psi \leftrightarrow (\varphi \rightarrow \neg[!\varphi]\psi)$.
(3) $[!\varphi](\psi \wedge \chi) \leftrightarrow ([!\varphi]\psi \wedge [!\varphi]\chi)$.
(4) $[!\varphi]K_a\psi \leftrightarrow (\varphi \rightarrow K_a[!\varphi]\psi)$.

As for inference rules, we have *Generalization* for the operator $[!\varphi]$, as well as *Replacement of Equivalents*.

Example 2.10 (soundness of reduction axioms) We do the crucial case of knowledge after announcement. This compares two models: $(\mathfrak{M}, s)$ and $(\mathfrak{M}|\varphi, s)$ before and after the update. It helps to draw pictures relating these to understand the following proof. The formula $[!\varphi]K_a\psi$ says that, in $\mathfrak{M}|\varphi$, all worlds $\sim_a$-accessible from s satisfy ψ. The corresponding worlds in $\mathfrak{M}$ are those worlds which are $\sim_a$-accessible from s *and which satisfy* ψ. Moreover, given that truth values of formulas may change in an update step, the correct description of these worlds in $\mathfrak{M}$ is not that they satisfy φ (which they do in $\mathfrak{M}|\varphi$), but rather $[!\varphi]\psi$: they become ψ after the update. Finally, $!\varphi$ is a partial operation, as φ has to be true for its public announcement. Thus, we need to make our assertion on the right conditional on $!\varphi$ being executable, i.e., φ being true. Putting all this together, $[!\varphi]K_a\psi$ says the same as $\varphi \rightarrow K_a(\varphi \rightarrow [!\varphi]\psi)$. But given the effect of the operator $[!\varphi]$ for a partial operation, we can simplify this final formula to the equivalent $\varphi \rightarrow K_a[!\varphi]\psi$.

At this point, readers may have felt that we should have added an axiom to deal with stacked modalities. The relevant validity of *PAL* is this:

$$[!\varphi][!\psi]\chi \leftrightarrow [!(\varphi \wedge [!\varphi]\psi)]\chi.$$

However, perhaps in contrast with the reader's expectations, such a principle is not required for a complete axiomatization. The reason is that we can always start with innermost dynamic modalities in given formulas, and reduce these out. At no stage is there a need to reduce two stacked dynamic modalities immediately.[4]

Note that the dynamic "reduction axioms" take every formula of our dynamic-epistemic language eventually to an equivalent formula inside the static pure epistemic language. Regarding logic, this means that *PAL* is complete, since the static epistemic logic is complete, and *PAL* is *decidable*, since this is true for its static epistemic base language.

[4] I have benefited from a discussion of this point with Johan van Benthem. His new book [32] contains a further analysis of this issue in terms of logical validities for various dynamic actions, and concrete illustrations in the context of games and protocol update.

2.4 Dynamic Epistemic Logic

Dynamic epistemic logic is a generalization of the public announcement logic, dealing with more complex communications. Typically, there are uncertainties concerning the events, agents do not know whether some event has happened or not. That may be resulted from the following two cases: Either there was some problem with the event itself, or there were some constraints on the receiver's capacity. Consider the following scenario:

Example 2.11 (two agents case) For an illustration, consider two students a and b who were present when Professor Zhang said: "The homework for today is Exercise 5 on Page 321." Student a heard what Professor Zhang said (call it p) because she was sitting in front. But student b did not, as he was sitting a bit further to the back: the professor might have said p, or the opposite $\neg p$.

Before the announcement, the epistemic situation can be pictured as (Fig. 2.3):

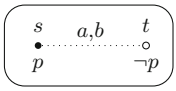


Fig. 2.3 Both agents do not know

Note that this is similar to the scenario depicted in Fig. 2.1 except that there are two agents a and b now.

Turning now to the epistemic status of the two agents regarding the event taking place, it can be modeled in the following diagram:

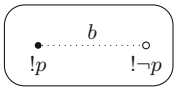


Fig. 2.4 Agent b does not hear precisely

As in our ordinary epistemic models, this model contains two possible events, "announcing p" and "announcing $\neg p$", with an uncertainty relation between them for one of the agents. "Announcing p" is what actually happened. In this case, only agent b is uncertain whether p or $\neg p$ was announced.

Next, we want to know the epistemic status of agent a and b after this event has happened. The *product update* we are going to introduce below gives us a general mechanism to handle such scenarios. It was first proposed in [14]. What it models are scenarios with possibly different observations by different agents, leading to differences in information flow, and forms of privacy. The most creative idea was to treat possible events as something similar to possible worlds, modeling agents' observation of the events in terms of standard epistemic uncertainty relations, as shown in Fig. 2.4. A second important idea is the fact that events come with *preconditions* for their occurrence, since this is what makes their observation informative.

Definition 2.12 (event models) An event model is a tuple $\mathcal{E} = (E, \sim_a, PRE)$ such that E is a non-empty set of events, $\sim_a$ 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 event: An event e can be performed at world s only if the world s fulfills the precondition $PRE(e)$.

Definition 2.13 (product update) Let an epistemic model $\mathfrak{M} = (S, \sim_a, V)$ and an event model $\mathcal{E} = (E, \sim_a, PRE)$ be given, the product update model is defined to be the model $\mathfrak{M} \otimes \mathcal{E} = (S \otimes E, \sim'_a, V')$ such as

(1) $S \otimes E = \{(s, e) \in S \times E : (\mathfrak{M}, s) \models PRE(e)\}$.
(2) $(s, e) \sim'_a (t, f)$ iff both $s \sim_a t$ and $e \sim_a f$.
(3) $V'(p) = \{(s, e) \in S \otimes E : s \in V(p)\}$.

As we can see from the definition, the product update model consists of all updated worlds in which events have taken place. The new uncertainty relation is determined by the previous relation between possible worlds and that between possible events. The valuation for atomic propositions remains the same as in the old worlds.

Let us return to Example 2.11 now. According to the definition of product update, we obtain the following updated model (Fig. 2.5):

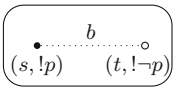


Fig. 2.5 Agent b still does not know

Inspecting this model, b did not hear what Professor said and does not know about the homework, but agent a heard and got to know. This fits our intuition precisely.

Though the product update mechanism is very simple, it can be widely applied to more complex scenarios, including private announcements in subgroups, security of information channels, and games of imperfect information. For more details of its technique and its scope, we refer the reader to [14] and [65].

The above notions suggests an extension of the epistemic language, defined in the following.

Definition 2.14 (dynamic epistemic language) Let a set of proposition variables Φ, a finite set of agents N, a set of events E be given. The dynamic epistemic language is defined by the rule

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid [e]\varphi \quad \text{where } p \in \Phi, a \in N, \text{and } e \in E.$$

The intended reading of formula $[e]\varphi$ is "after event e takes place, φ holds." We could also add the usual action operations of *composition*, *choice*, and *iteration* from propositional dynamic logic to the event vocabulary, so that we can talk about more complex events. But in this chapter, we will have no special use for these. The

language has new dynamic modalities $[e]$ referring to epistemic events, and these are interpreted in the product update model as follows[5]:

$$\mathfrak{M}, s \models [e]\varphi \quad \text{iff} \quad \text{if } \mathfrak{M}, s \models PRE_{(e)}, \text{ then } \mathfrak{M} \otimes \mathcal{E}, (s, e) \models \varphi.$$

There is a complete axiomatization for the dynamic logic of product update in the given language. Instead of describing this logic in detail, we merely state its most important reduction axiom, the one concerning knowledge. The following valid equivalences encodes how knowledge changes when complex informational events take place:

$$[e]K_a\varphi \leftrightarrow PRE(e) \rightarrow \bigwedge_{f \in E}\{K_a[f]\varphi : e \sim_a f\}.$$

Intuitively, saying that 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.

Both the basic public announcement update and the more sophisticated product update for private information with complex triggering events can be adapted to handle preference or evaluation change. We will illustrate how this works step by step in the coming chapters.[6]

2.5 Methodology

There are several major features to the *DEL* approach, which make it a general methodology.[7] So far, it has been applied in analyzing knowledge update, information-driven changes in beliefs (cf. [16, 29]), changes in intentions [164], and in preference: the main theme of this book ([83, 131]). Very recent applications include the information dynamics of acts of inference [191] and of questions as acts of issue management (cf. [43]).

Of course such applications often come with new twists. For instance, and very importantly to what follows, update in beliefs or preferences cannot be an eliminative model change in the public announcement style. Instead, it leaves the universe of worlds the same, but it *changes the ordering pattern* through a systematic change in the current relations between possible worlds. We will discuss some more general methodological features of *DEL* here, to conclude this chapter, and prepare the reader for what is to come.

---

[5] Note that the pre-conditions in the event model are formalized in the dynamic epistemic language, this might introduce circularities. One solution is to "view the preconditions as literal parts of event models". We refer to [32] and [65] for further dicussions.

[6] In this book, we will not use the original privacy motivations for product update. In our later applications to qualitative and quantitative preference change, the different events are rather different "signals" that can apply to worlds, whose ordering tells us something about how agents are to evaluate their relative betterness or plausibility.

[7] For a first programmatic discussion in this vein, see [29].

First, given a static epistemic logic, *DEL* "dynamifies" it and adds on a dynamical superstructure. This dynamification is, in principle, independent from the base system – we can choose any static system we like. But there are constraints here. Whatever modal base logic we have chosen, its matching frame properties should be kept after the update.[8] This pattern of dynamification works well across a wide range of logics. Indeed, *DEL* may be seen as a way of "upgrading" the operating system of existing philosophical logic to much higher functionality.

Secondly, looking ahead, *DEL* describes what agents would know after some dynamic event has happened. The knowledge change before and after the event are characterized in so called *reduction axioms*. Reduction axioms are equivalences taking each formula of our dynamic epistemic language eventually to an equivalent formula in the static epistemic base language. As a result, if the static base logic is complete or decidable, its dynamic extension is complete or decidable, too. The earlier-mentioned upgrade comes for free.[9]

Thirdly, in applications of the paradigm, sometimes, we cannot find reduction axioms for certain operators in our language. A typical example is *common knowledge*, for which the earlier logic *PAL* has no reduction axiom. To solve this problem, one must enrich the base language. For instance, one can introduce a new notion of *conditional common knowledge* in the static language in order to find a *PAL* reduction axiom for common knowledge.[10]

Finally, recent versions of *PAL* and *DEL* have back-pedaled a bit on the strict reduction methodology. The reason is that the reduction axioms as stated have one presupposition that may not always be natural: in checking their soundness, one uses the fact that the action *is always available*. But many natural scenarios of information flow have procedures restricting available informational events. Civilized conversation need not allow all that is true to also be said, a feasible medical procedure only allows for certain test on patients in certain orders, etc. To model this, one needs a longer-term temporal perspective, with a *protocol* regulating the sequences of successive informational events that can occur. As a result, dynamic logics now acquire a temporal element with less drastic reduction axioms, though they can still be axiomatized completely. See [35] for this temporal-style *DEL* and its relations to epistemic-temporal logics of branching time. Reference [110] develops its underlying philosophical idea of "procedural information" in much more detail, adding new applications to epistemology.

In the next part of this book, we will start our investigation of our main themes: preference representation and preference dynamics. A simple modal model for preference will be introduced, and we will explore how to model events where new

[8] Say, with epistemic **S5**, the frame still has equivalence relations after *DEL* update.

[9] Still, the reduction does not settle computational complexity: Translation via the axioms may increase formula length exponentially. Still, for the case of *PAL*, the complexity of satisfiability remains that of epistemic logic, viz. Pspace-complete (cf. [138]).

[10] This again requires a reduction axiom for conditional common knowledge, which can be done as well. A good reference for such issues of language design in *DEL* is [34].

information changes our preference, but also where preference changes occur that are *sui generis*, such as commands. While the setting will be different, the reader will recognize many of the ideas and techniques of this chapter in the following chapters. She will hopefully agree that what worked for information dynamics, also makes a lot of sense in this new area of agents' evaluation of the worlds.

Part III
Preference over Worlds

Chapter 3
Preference over Worlds: Static Logic

3.1 Introduction

Preferences arise from comparisons between alternatives, say, outcomes, actions, or situations. Such a comparison is typically associated with some ordering, indicating that one alternative is "better" than another. For instance, when playing chess or other games, choosing a move π_1 instead of π_2 is determined largely by a consideration concerning the outcomes that π_1 or π_2 leads to. In general, individual preferences can be used to predict behavior by rational agents, as studied in game theory and decision theory. Preference logics in the literature study the abstract properties of different comparative structures [101].

Preference statements can be weaker or stronger in what they say about alternatives 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 a somewhat objective approach, where a binary primitive preference relation in possible worlds models supports a unary modality "true in some world which is at least as good as the current one" (similar models have been studied in [55] and [94]). We will call this relation "betterness" to distinguish it from richer notions of preference. Then we will use a standard modal language to express preference and its related notions. We will show that this language is very expressive, being able to express various kinds of preference that agents may have between propositions, i.e., types of events.[1]

This chapter is structured as follows. In Section 3.2 we will introduce the language and semantics for modal betterness logic. A complete axiomatization will be stated. In Section 3.3 we will study expressive power, with a special interest in defining preference over propositions ("generic preference") in terms of *lifting* the primitive betterness relation from possible worlds to an ordering over propositions, viewed as sets of possible worlds. Various possible lifts will be considered. Finally, as an illustration, Section 3.4 will focus on one type of lift, namely the $\forall\exists$-version,

[1] In a different setting, [44] showed how such a language, extended with hybrid modalities, defines conditionals, Nash equilibrium, and Backward Induction solutions to games.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_3,

and explore its logical properties in more detail. Our conclusion summarizes the platform laid in this chapter for the rest of this book.

3.2 Modal Betterness Logic

As we commented in the above, semantical betterness relations are treated as modalities in the language. We now introduce the language of modal betterness logic:

Definition 3.1 (modal betterness language) Take any set of propositional variables Φ, with a variable p ranging over Φ. The modal betterness language $\mathcal{L}_{\mathcal{B}}$ is given by the following inductive syntax rule:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \langle\leq\rangle\varphi \mid \langle<\rangle\varphi \mid E\varphi.$$

The intended reading of $\langle\leq\rangle\varphi$ is "φ is true in some world that is at least as good as the current world", while $\langle<\rangle\varphi$ says that "φ is true in some world that is strictly better than the current world."[2] We will discuss how these two modalities help lift betterness relations to preferences over propositions in Section 3.3, in particular, in the case where the betterness relations lack connectedness. In general, these notions are agent-relative, but in what follows we will mostly suppress this aspect, since it is orthogonal to our main points. In addition, the auxiliary existential modality $E\varphi$ says that "there is some world where φ is true". Combinations of these modalities can capture a wide variety of binary preference statements comparing propositions, again, we will show this soon.

As usual, we will write $[\leq]\varphi$ for the universal modality $\neg\langle\leq\rangle\neg\varphi$, and we will write $[<]\varphi$ and $U\varphi$ for the duals of $\langle<\rangle\varphi$ and E, respectively. Either $[\leq]$ or $\langle\leq\rangle$ can be introduced as a primitive, we will take the technical convenience into account and use both formats interchangeably in this context.

How is this formal language connected to "preference" as it occurs in natural discourse? One may be inclined to read $\langle\leq\rangle\varphi$ as "some agent 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 $[\leq]\varphi$ "in all better worlds" seems much too strong. Cf. [102] 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 feature. Our approach emphasizes comparisons of worlds, rather than propositions, whereas common notions of preference often play between propositions, or semantically, sets of worlds. Even so, the preferences between propositions can be *defined* in the present language, as a lift from the

[2] We use two independent modalities here for weak and strict betterness. This may seem strange, since strict order was definable in terms of weak order. But the point is that this definition cannot be reproduced in a natural way inside our modal language, which therefore brings out reasoning with both modalities on a par.

betterness relations. We will study those lifts soon in Section 3.3. For the moment, we just take this expressive power of our modal language for granted. The virtue of working with simple base modalities, as we do, is that these "decompose" many more complex preference statements in a perspicuous manner, while allowing for a simple dynamic approach later on.

Definition 3.2 (modal betterness model) A modal betterness model is a tuple $\mathfrak{M} = (S, \leq, V)$ where S is a set of possible worlds, $\leq$ is a reflexive and transitive relation (the 'betterness' pre-order) over these worlds, and V is a valuation assigning truth values to proposition letters at worlds.[3]

We read $s \leq t$ as "t is at least as good as s", or "t is weakly better than s". If $s \leq t$ but not $t \leq s$, then t is *strictly better* than s, written as $s < t$. If $s \leq t$ and $t \leq s$, then s and t are *indifferent*.

Note that we do not require that our betterness relations be *connected* in the sense of the Lewis sphere models for conditional logic. In general, we want to 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. It is a very natural situation we may often encounter in real life. 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.

Definition 3.3 (truth conditions) Given a modal betterness model $\mathfrak{M} = (S, \leq, V)$, and a world $s \in S$, we define $\mathfrak{M}, s \models \varphi$ (formula φ is true in $\mathfrak{M}$ at s) in the usual manner by induction on the construction of the formula φ:

$\mathfrak{M}, s \models \top$ iff always.

$\mathfrak{M}, s \models p$ iff $s \in V(p)$.

$\mathfrak{M}, s \models \neg\varphi$ iff not $\mathfrak{M}, s \models \varphi$.

$\mathfrak{M}, s \models \varphi \wedge \psi$ iff $\mathfrak{M}, s \models \varphi$ and $\mathfrak{M}, s \models \psi$.

$\mathfrak{M}, s \models \langle\leq\rangle\varphi$ iff for some t with $s \leq t$, $\mathfrak{M}, t \models \varphi$.

$\mathfrak{M}, s \models \langle<\rangle\varphi$ iff for some t with $s < t$, $\mathfrak{M}, t \models \varphi$.

$\mathfrak{M}, s \models E\varphi$ iff for some world t in S, $\mathfrak{M}, t \models \varphi$.

Definition 3.4 (modal equivalence) Two models $\mathfrak{M}, s$ and $\mathfrak{M}', t$ are modally equivalent, written as $\mathfrak{M}, s \leftrightsquigarrow \mathfrak{M}', t$ if they satisfy the same formulas from $\mathcal{L}_{\mathcal{B}}$.

---

[3] In this chapter, we use pre-orders since we want the generality of possibly non-total preferences. Total orders, the norm in areas like game theory, provide an interesting specialization for the results in this chapter. We will study total ordered preference in Chapter 7.

Definition 3.5 (bisimulation) Two models $\mathfrak{M}, s$ and $\mathfrak{M}', t$ are bisimilar (written $\mathfrak{M}, s \triangleq \mathfrak{M}', t$) if there is a relation $Z \subseteq S \times S'$ such that:

(1) If sZt then for all $p \in \Phi$, $s \in V(p)$ iff $t \in V(p)$.
(2) If sZt and $s \leq s'(s < s')$ then there is a $t' \in S'$ such that $t \leq t'(t < t')$ and $s'Zt'$. (the forth condition).
(3) If sZt and $t \leq t'(t < t')$ then there is a $s' \in S$ such that $s \leq s'(s < s')$ and $s'Zt'$. (the back condition).
(4) For all $s \in S$, there is a $t \in W'$ such that sZt.
(5) For all $t \in S'$, there is a $s \in W$ such that sZt.

This is what is called a total bisimulation, as it includes conditions 4 and 5. It is easy to show that any two bisimular models are modally equivalent with regard to our language, in other words, we say that the language is bisimulation-invariance. Bisimular models are often used to show undefinability of certain operators in some language. We will come back to this issue in Section 3.3.

As for the resulting logics, [39] give a complete axiomatization for the logic of weak and strict betterness modalities. Essentially, the system consists of **S4**-axioms for operator $[\leq]$, **K** for $[<]$, and **S5**-axioms for universal modality U, plus some axioms for interaction between operators. We restate it here, omitting the proof.

Theorem 3.6 *The modal betterness logic is completely axiomatized by the following set of principles:*

(1) *propositional tautologies*
(2) $[\leq](\varphi \to \psi) \to ([\leq]\varphi \to [\leq]\psi)$
(3) $[<](\varphi \to \psi) \to ([<]\varphi \to [<]\psi)$
(4) $[\leq]\varphi \to \varphi$
(5) $[\leq]\varphi \to [\leq][\leq]\varphi$
(6) $U(\varphi \to \psi) \to (U\varphi \to U\psi)$
(7) $U\varphi \to \varphi$
(8) $U\varphi \to UU\varphi$
(9) $\neg U\varphi \to U\neg U\varphi$
(10) $[\leq]\varphi \to [<]\varphi$
(11) $[<]\varphi \to [\leq][<]\varphi$
(12) $[<]\varphi \to [<][\leq]\varphi$
(13) $[\leq]([\leq]\varphi \vee \psi) \wedge [<]\psi \to \varphi \vee [\leq]\psi$
(14) $U\varphi \to [\leq]\varphi$
(15) $E\varphi \leftrightarrow \neg U\neg\varphi$
(16) $\langle\leq\rangle\varphi \leftrightarrow \neg[\leq]\neg\varphi$
(17) $\langle<\rangle\varphi \leftrightarrow \neg[<]\neg\varphi$

The correspondence between the above axioms and the frame properties can be studied just as one does in standard modal logics [24]. Additional axioms in our language impose further frame conditions on models. Nevertheless, we will work with the minimal system described above, leaving such extras aside.

3.3 Expressive Power

Our modal base language seems so simple and standard that it may be hard to see what it can define by way of more complex notions relevant to preference. In this section, we will show that it can express more than the reader might have thought. We will give two examples: "generic preferences" and "conditional preference".

Preference between specific worlds, introduced above, is just as in decision theory and game theory. But preference can be used to compare different sorts of things. In game theory, we do need to compare kinds of situation. Starting with von Wright, logicians have studied "generic preferences" between kinds of object, or kinds of situation. Such scenarios, too, occur in many other fields in the literature, with various interpretations of the basic relation $y \leq x$. It is interpreted as "x is as least as normal (or typical) as y" in [55] on conditional and default reasoning, as "x at least as preferred or desirable as y" in [66], as "x is no more remote from actuality than y" in [127] on counterfactuals, and as "x is as likely as y" in [94] on qualitative reasoning with probability. In all these settings, it makes sense to extend the given order on worlds to an order of propositions φ, ψ. 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 [21].

In what follows, we will show that preference over propositions is definable as a binary operator $P(\varphi, \psi)$ in our modal logic, whose language is rich enough to explicitly define "lifts" of betterness on worlds to a binary ordering on sets of worlds. Studies of such lifts abound (cf. [42, 44] and [39]), and our purpose in this section is merely to streamline some results.

3.3.1 Generic Preference: Quantifier Lifts

One obvious way of lifting world orders $x \leq y$ to proposition or set orders $X \trianglelefteq Y$ uses definitional schemas that can be classified by the quantifiers which they involve. As has been observed by many authors (cf. [39]), there are four obvious two-quantifier combinations for lifting:

(1) $X \trianglelefteq^{\forall\forall} Y \Leftrightarrow \forall x \in X\, \forall y \in Y\colon x \leq y$;

(2) $X \trianglelefteq^{\forall\exists} Y \Leftrightarrow \forall x \in X\, \exists y \in Y\colon x \leq y$;

(3) $X \trianglelefteq^{\exists\forall} Y \Leftrightarrow \exists x \in X\, \forall y \in Y\colon x \leq y$;

(4) $X \trianglelefteq^{\exists\exists} Y \Leftrightarrow \exists x \in X\, \exists y \in Y\colon x \leq y$.

Taking the strict version of the betterness relation gives four more combinations:

(5) $X \lhd^{\forall\forall} Y \Leftrightarrow \forall x \in X\, \forall y \in Y: x < y$;

(6) $X \lhd^{\forall\exists} Y \Leftrightarrow \forall x \in X\, \exists y \in Y: x < y$;

(7) $X \lhd^{\exists\forall} Y \Leftrightarrow \exists x \in X\, \forall y \in Y: x < y$;

(8) $X \lhd^{\exists\exists} Y \Leftrightarrow \exists x \in X\, \exists y \in Y: x < y$.

As usual, we can define $X \lhd Y$ as $X \trianglelefteq Y$ and $\neg Y \trianglelefteq X$. One can argue for any of these as a notion of generic preference. Reference [44] claims that $\lhd^{\forall\forall}$ is the notion of "preference" intended by von Wright in his seminal work on preference logic [197] and provides an axiomatization. But the tradition is much older, and (modal) logics for preference relations over sets of possible worlds have been considered by [55, 127] and [94], and other authors. In particular, [94] studied the above $\forall\exists$-combination.

We are not in the position to claim that one lift is more plausible than another, but our main concern here is the logical properties of lifts, and the expressive power of our modal betterness language.

3.3.2 Expressing Generic Preferences in $\mathcal{L}_{\mathcal{B}}$

So far, our discussion of preference over propositions has been semantic-oriented. A natural question is the following: can we express the above generic preferences in the language $\mathcal{L}_{\mathcal{B}}$? The answer is positive. We start with the following four:

Definition 3.7 (**generic preference: $\forall\exists$ and $\exists\exists$**) The $\forall\exists$-preference and $\exists\exists$-preference can be defined in the language $\mathcal{L}_{\mathcal{B}}$ as follows:

(1) $\varphi \trianglelefteq^{\forall\exists} \psi := U(\varphi \to \langle\leq\rangle\psi)$

(2) $\varphi \trianglelefteq^{\exists\exists} \psi := E(\varphi \wedge \langle\leq\rangle\psi)$

(3) $\varphi \lhd^{\forall\exists} \psi := U(\varphi \to \langle<\rangle\psi)$

(4) $\varphi \lhd^{\exists\exists} \psi := E(\varphi \wedge \langle<\rangle\psi)$

We can read $\varphi \trianglelefteq^{\forall\exists} \psi$ as "for each φ-world, there exists a ψ-world which is as good as that φ-world", and read $\varphi \lhd^{\forall\exists} \psi$ as "for each φ-world, there exists a better ψ-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, [94] 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 S 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 [127], where the given order on S is quasi-linear.

Now let us consider the remaining combinations. It turns out to be more complex and interesting, having to do with the models under discussion in the following sense. If the betterness relations satisfy the property of *connectedness*, we can make use of the Definition 3.7 and define those cases as follows:

Definition 3.8 (**generic preference: ∀∀ and ∃∀**) The ∀∀-preference and also the ∃∀-preference can be defined in the language $\mathcal{L}_{\mathcal{B}}$ on connected models:

(1) $\varphi \trianglelefteq^{\forall\forall} \psi := U(\psi \rightarrow [<]\neg\varphi)$

(2) $\varphi \trianglerighteq^{\exists\forall} \psi := E(\varphi \wedge [<]\neg\psi)$

(3) $\varphi \triangleleft^{\forall\forall} \psi := U(\psi \rightarrow [\leq]\neg\varphi)$

(4) $\varphi \triangleright^{\exists\forall} \psi := E(\varphi \wedge [\leq]\neg\psi)$

However, if we drop connectedness, basic modal definability fails:

Fact 3.9 *The ∀∀ and ∃∀-meaning of propositional preference cannot be defined in the modal betterness language $\mathcal{L}_B$ on non-connected models.*

Proof This can be shown by a standard argument, providing two bisimilar models one of which satisfies the lift, and one of which does not (cf. [44]). □

Thus, some lifts would amount to genuine first-order extensions of the modal base language, in the spirit of hybrid logics. While we will not pursue such extensions in our book, many of the results that we develop will survive such generalizations.[4]

3.3.3 Conditional Preference

Here is one more example that shows the expressive power of our basic language. A widespread *maximality operator* $[Best(\psi)]\varphi$ in the literature on conditional or deontic logic says that the "best' ψ-worlds in some relevant order satisfy some proposition φ". The counterpart of this notion may be called *conditional preference*, and it can be expressed as follows in our language:

(1) $$P^{\psi}\varphi := U(\psi \rightarrow \langle\leq\rangle(\psi \wedge [\leq](\psi \rightarrow \varphi))).$$

This says that φ is preferred on condition of ψ, if and only if, for all worlds that satisfy ψ, there is a better ψ-state such that all ψ-states above it are φ. A similar modal definition for conditionals was first proposed by [53] and [55]. Here, we

[4] One example are the added "intersection modalities" of Chapter 5.

restated it in our language. We omit the simple verification that this formula really has the intended meaning, at least on finite models.

Interestingly, our base language even offers another definition, whose syntax is even closer to the maximality clause for the antecedent. For instance, using the strict betterness modality, [83, Ch. 3] defines conditional preference as

$$P^{\psi}\varphi \;:=\; U((\psi \wedge \neg\langle < \rangle \psi) \to \varphi).$$

A few comments are in order here. First, for these definitions to capture the intended meaning of the maximality operator, finiteness, or more generally, converse well-foundedness of the ordering should be assumed, to make sure that maximal worlds exist satisfying given formulae. In the absence of converse well-foundedness, as was observed in [55], Formula (1) expresses something more than maximality.

Next, what we showed here does not just apply to betterness and preference. Following the same idea, given an order relation of "relative plausibility", we can define *conditional beliefs*, which will be important in many of our later chapters. Formula (1) above then amounts to the standard syntactic "relativization" of absolute belief (as truth in all most plausible worlds) to just the worlds satisfying the antecedent. Moreover, we will often relativize this still further, restricting everything to just the set of worlds that are *epistemically accessible* from the current one. Thus, in Chapter 5, we will use a knowledge modality K instead of the universal modality U when lifting "epistemically entangled" betterness relations, while dealing analogously with plausibility relations.

For now, we just emphasize what all this has shown. Our basic modal betterness logic can encode quite a few complex properties of preference – as well as, reinterpreting the basic world order, of related notions such as belief.

3.4 Preservation and Characterization of ∀∃-Preference

In this section, mainly an excursion, we will raise a few more semantic issues, to further understanding the lifting phenomenon per se.

Among various kinds of lift, 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 $y \leq x$.

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, namely, ∀∃-preference, and try to understand better how the lift generally works. One question that comes to mind immediately is this: Can the properties of 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 Section 3.2. Can we show $\trianglelefteq^{\forall\exists}(\varphi, \psi)$ has these two properties? In fact, the answer is positive. We can even prove something stronger:

Fact 3.10 *Reflexivity and transitivity of the relation $\leq$ are preserved in the lifted relation $\trianglelefteq^{\forall\exists}$, but also vice versa.*

Proof Reflexivity. To show that $\trianglelefteq^{\forall\exists}(X, X)$, by Definition 3.7, we need that $\forall x \in X \exists y \in X : x \leq y$. Since we have $x \leq x$, take y to be x, and we get the result.

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

Transitivity. Assume that $\trianglelefteq^{\forall\exists}(X, Y)$ and $\trianglelefteq^{\forall\exists}(Y, Z)$. We show that $\trianglelefteq^{\forall\exists}(X, Z)$. By Definition 3.7, this means we have $\forall x \in X \exists y \in U : x \leq y$ and $\forall y \in Y \exists z \in Z : y \leq z$. Then by transitivity of the base relation, we have that $\forall x \in X \exists z (x \leq z$, and this is precisely $\trianglelefteq^{\forall\exists}(X, Z)$.

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

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

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

Next, staying at the level of propositions, suppose we have a preference relation that is a ∀∃-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:

Theorem 3.11 (characterization) *A binary relation $\trianglelefteq$ over propositions satisfies the following four properties iff it is a ∀∃-lifting of some preference relation over the underlying possible worlds.*

1. $Y \trianglelefteq X \Rightarrow Y \bigcap Z \trianglelefteq X$ (*left downward monotonicity*)
2. $Y \trianglelefteq X \Rightarrow Y \trianglelefteq X \bigcup Z$ (*right upward monotonicity*)
3. $\forall i \in I, Y_i \trianglelefteq X \Rightarrow \bigcup_i Y_i \trianglelefteq X.$ (*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 ∀∃-lifting. We show that $\trianglelefteq$ has the four properties.

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

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

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

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

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

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

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 ($\dagger$), $y \leq x$. This shows that, for any $y \in Y$, there exists some $x \in X$ s.t. $y \leq x$, which is to say that $Y \trianglelefteq^{\forall\exists} X$.

($\Leftarrow$) Assume that $\forall y \in Y \exists x \in X : y \leq x$. By Definition ($\dagger$), $y \leq x$ 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 $Y \trianglelefteq X$. □

We have spent so much time with this one generic preference because it occurs quite widely in the literature, providing a typical use of our modal preference logic. But of course, many other lifts could be analyzed in a similar manner.

3.5 Conclusion

We have introduced a static modal betterness language in this chapter, which admits of a complete axiomatization for its valid forms of reasoning with preference. We then showed that the betterness relation over possible worlds can be lifted to generic preferences over propositions using various quantifier combinations. We discussed different lifts, giving us a better impression of the expressive power and limitations of our modal language. Finally, we characterized one lift completely, the widespread and well-behaved notion of "$\forall\exists$-preference".

This modal betterness language is just a launching platform in this book. As we will see in the next chapter, it provides a natural model for the dynamics of preference change, both at the basic level of modification in the betterness relations, and derived from that, at the level of defined "lifted" generic preferences between propositions.

Chapter 4
Preference over Worlds: Dynamic Logic

4.1 Introduction

In the preceding chapter, we have introduced a logical framework for static preference. Continuing on this platform, our main concern in this chapter is the *dynamics* of preference change. Our preferences are modified constantly through commands of moral authorities, suggestions from friends who give good advice, or just changes in our own evaluation of worlds and actions. Living in a society, our preference is also affected by what others like or dislike, as vividly described in the Chinese classic *Record of Music*[1]:

> *A ruler has only to be careful of what he likes and dislikes. What the ruler likes, his ministers will practise; and whatever superiors do, their inferiors will follow.*

Thus preference changes can have various triggers. In this chapter, we will concentrate on two kinds of trigger: informational events, and immediate betterness changing events. We start with the latter, since they reveal more of what is intrinsic to preference. Consider the following example:

Example 4.1 (taking a trip) Alice is indifferent between taking a trip (written as p) and staying at home ($\neg p$). Now her friend Bob comes along and says

"Let's take a trip!"

Alice's "taking" this *suggestion* means that any preference she might have had for staying at home is removed from the current model.

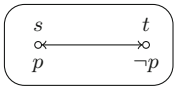
This dynamic process is pictured below. Arrows point at equally or more preferred worlds.[2] We can see from Fig. 4.1 that the two worlds s and t are indifferent in the initial model.

[1] The is the first work on music in Chinese history. It is believed that the book was written about two thousands year ago, but the precise time and author are still controversial.

[2] In addition to the arrows drawn, our betterness relations always have reflexive loops.

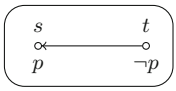
F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_4, © Springer Science+Business Media B.V. 2011

Fig. 4.1 Initial model



But in the upgraded model, the world s has become better, and hence preferable:

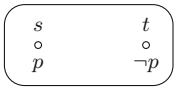
Fig. 4.2 Upgraded model



Thus, in our scenario, a suggestion removes already existing preference links: But it does not add new ones. This simple mechanism will be studied in greater detail in Section 4.3, as a point of departure for more general kinds of betterness upgrade. For the moment, by way of contrast, here is one more example, which does not remove links, but rather adds them.

In Fig. 4.2 above, Alice now prefers the trip after Bob's suggestion , 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 p does not necessarily 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 at the first stage, as in the following model with two unrelated worlds (Fig. 4.3):

Fig. 4.3 A model with two unrelated worlds



A suggestion in the above relation-decreasing sense does not make the worlds comparable. This is different with real commands like "Take that trip!", maybe coming from Alice's boss Carol who thinks it is necessary for her to take a rest. In such a case, we want to make sure Alice now prefers p. Then, we need to *add* preference links to the picture, making the world with $\neg p$ less preferred. Our later proposals in this chapter also deal with upgrades that add links between worlds.

As for other relevant scenarios motivating our study, consider a process of *planning*. We start with just our own initial goals in mind, and 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 studied by many authors, including [33, 100, 185, 200], and [199].

Related ideas of preference change play in dynamic semantics for conditional logics (for instance, in [176] and [192]). In its static Lewis-style semantics, a conditional $\varphi \Rightarrow \psi$ says roughly the following about the current model:

ψ is true in all most-preferred φ-worlds (♮)

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. Or even more simply than this, consider a default assertion

"*Normally* φ."

As [192] 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.

The aim of this chapter is to provide a logical format for studying preference dynamics in its appropriate generality. In Section 4.2, we start from a simple test scenario called a "suggestion", analyzing its induced betterness relation change, and determining its complete dynamic betterness logic. We will also look at how this systematically changes the generic preferences studied in Chapter 3. Section 4.3 will then generalize these results to a much wider family of patterns for dynamic relation change. As a final illustration, Section 4.4 will show how the resulting framework can also be applied to doxastic reasoning with defaults. We end up with some discussion and general conclusions.

4.2 Dynamic Betterness Logic

4.2.1 Upgrade as Relation Change

With the paradigm of public announcement in mind, we consider Example 4.1 once more, and define the mechanism of betterness change. Now, our static models are of course the modal betterness relational structures of Section 3.2:

$\mathfrak{M} = (S, \leq, V)$

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

$\sharp\varphi$

Intuitively, acts of "taking a suggestion" lead to the following model change, removing preferences for $\neg\varphi$ over φ:

Definition 4.2 (upgraded model) Given any modal betterness model $(\mathfrak{M}, s)$, the upgraded model $(\mathfrak{M}_{\sharp\varphi}, s)$ is defined as follows:

(1) $(\mathfrak{M}_{\sharp\varphi}, s)$ has the same domain, valuation, and actual world as $(\mathfrak{M}, s)$, but
(2) the new betterness relations are now

$$\leq^{*} = \leq - \{(s, t) \mid \mathfrak{M}, s \models \varphi \text{ and } \mathfrak{M}, t \models \neg\varphi\}.^{3}$$

Next, we define the new dynamic language formally below, it is an extension of the static modal betterness language:

[3] Reference [105] analyzes newly defined betterness relations in a set-theoretic format.

Definition 4.3 (dynamic betterness language) Take any set of propositional variables Φ, with a variable p ranging over Φ. The dynamic betterness language is given by a mutually recursive syntax rule:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \langle\leq\rangle\varphi \mid \langle<\rangle\varphi \mid E\varphi \mid [\pi]\varphi$$
$$\pi := \sharp\varphi.$$

Here, the intended reading of the central new formula $[\sharp\varphi]\psi$ is that "after φ is publicly suggested, ψ holds".

Definition 4.4 Given a modal betterness model $\mathfrak{M}$, the truth definition for formulas is as before, but with one new key clause for the action modality:

$$(\mathfrak{M}, s) \models [\sharp\varphi]\psi \quad \text{iff} \quad \mathfrak{M}_{\sharp\varphi}, s \models \psi.$$

Differently from what we have seen for public announcements in Chapter 2, there are no preconditions that need to be satisfied for actions of suggestion. This fits our intuitions: People can suggest ideas that might turn out to be wrong.

As for an axiomatization, we have the following completeness result:

Theorem 4.5 *Dynamic betterness logic is axiomatized by the following axioms*[4]*:*

(1) *All theorems of modal betterness logic.*
(2) $\langle\sharp\varphi\rangle p \leftrightarrow p$.
(3) $\langle\sharp\varphi\rangle\neg\psi \leftrightarrow \neg\langle\sharp\varphi\rangle\psi$.
(4) $\langle\sharp\varphi\rangle(\psi \wedge \chi) \leftrightarrow (\langle\sharp\varphi\rangle\psi \wedge \langle\sharp\varphi\rangle\chi)$.
(5) $\langle\sharp\varphi\rangle\langle\leq\rangle\psi \leftrightarrow (\neg\varphi \wedge \langle\leq\rangle\langle\sharp\varphi\rangle\psi) \vee (\langle\leq\rangle(\varphi \wedge \langle\sharp\varphi\rangle\psi))$.
(6) $\langle\sharp\varphi\rangle E\psi \leftrightarrow E\langle\sharp\varphi\rangle\psi$.

Proof This is a set of reduction principles for upgrade similar to those for public announcement in Chapter 2. Axiom 2 for atomic sentences is even simpler – as we have said in the above, 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 (Axiom 5) which is crucial, as it encodes precisely how we change the betterness relation. It says essentially this. After an upgrade for φ, a betterness 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 more succinctly in the two disjuncts on the right-hand side. Finally, as the upgrade may have changed truth values of

[4] Just as with actions of public announcement in Chapter 2, the completeness theorem here does not give us an explicit valid principle for dealing with iterated modalities $\langle\sharp\varphi\rangle\langle\sharp\psi\rangle\chi$. Indeed, there is no such principle, as the effect of two consecutive suggestions may be genuinely different from making just one suggestion. Consider a simpler principle $\langle\sharp\varphi\rangle\langle\sharp\varphi\rangle\psi \leftrightarrow \langle\sharp\varphi\rangle\psi$, it holds for factual assertions φ. But it need not hold for non-factual φ which themselves refer to the ordering. After the first $\sharp\varphi$ action, we have changed the ordering, and the worlds where φ is now true need not be the same ones as those where φ was true before. Thus there is more structure to our dynamic logics than what we have brought to light so far.

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. Axiom 6 is simply a commutation law for betterness and existential modalities. □

This result is not the end of dynamic betterness logic, but the beginning. The above style of thinking works for a whole family of completeness theorems for betterness-changing operators, as we will see in the next section. The key to this is choosing an appropriate abstraction level, namely an abstract reading of a suggestion $\sharp(\varphi)$ as a sort of *relational program*

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

The technical computation behind our crucial reduction axiom is then this:

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

Similar results hold for dynamic logics with many other strong and weak commands (cf. [29, 42] and [131]). More details on this will come later.

4.2.2 Dynamics of Generic Preferences

This dynamic betterness logic can explain general effects of changes in 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 are obtained after an upgrade action, and relate these to the propositional preferences that we had before. As an illustration, consider the earlier $\forall\exists$-notion of preference:

$$\psi \trianglelefteq^{\forall\exists} \varphi \quad \text{iff} \quad U(\psi \rightarrow \langle \leq \rangle \varphi).$$

Fact 4.6 *The following equivalence is provable in dynamic betterness logic:*

$$\langle \sharp A \rangle \psi \trianglelefteq^{\forall\exists} \varphi \quad \textit{iff} \quad (\langle \sharp A \rangle \psi) \trianglelefteq^{\forall\exists} (\langle \sharp A \rangle \varphi) \wedge (\langle \sharp A \rangle \psi \wedge A) \trianglelefteq^{\forall\exists} (\langle \sharp A \rangle \varphi \wedge A).$$

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

$$\begin{aligned}
&\langle \sharp A \rangle \psi \trianglelefteq^{\forall\exists} \varphi \leftrightarrow \langle \sharp A \rangle U(\psi \rightarrow \langle \leq \rangle \varphi) \\
&\leftrightarrow U(\langle \sharp A \rangle (\psi \rightarrow \langle \leq \rangle \varphi)) \\
&\leftrightarrow U(\langle \sharp A \rangle \psi \rightarrow \langle \sharp A \rangle \langle \leq \rangle \varphi) \\
&\leftrightarrow U(\langle \sharp A \rangle \psi \rightarrow (\neg A \wedge \langle \leq \rangle \langle \sharp A \rangle \varphi) \vee (\langle \leq \rangle (A \wedge \langle \sharp A \rangle \varphi)))
\end{aligned}$$

$$\leftrightarrow U(\langle\sharp A\rangle\psi \wedge \neg A \rightarrow \langle\leq\rangle\langle\sharp A\rangle\varphi) \wedge U(\langle\sharp A\rangle\psi \wedge A \rightarrow \langle\leq\rangle(\langle\sharp A\rangle\varphi \wedge A))$$
$$\leftrightarrow (\langle\sharp A\rangle\psi) \trianglelefteq^{\forall\exists} (\langle\sharp A\rangle\varphi) \wedge (\langle\sharp A\rangle\psi \wedge A) \trianglelefteq^{\forall\exists} (\langle\sharp A\rangle\varphi \wedge A).$$

□

A similar analysis applies to the other notions of generic preference considered in Chapter 3. The required calculations are simple exercises.

One might want to be more radical here, drop the reduction, and insist on dynamic preference-changing actions directly *at the level of propositions*, without any dependence on an underlying world-level, deriving a world-level change only afterwards. This is in line with syntactic versions of belief revision theory where one is instructed to come to believe certain propositions. We will study this second way of thinking in Chapter 7, and once more in Chapter 10.

4.3 Preservation and General Betterness Transformers

4.3.1 Preservation Properties of Upgrade

Perhaps the most pressing issue in adding a dynamic update component to a static base logic 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 betterness relations. This time, no general result comes to the rescue, since we only have the following counterpart to the preservation result for submodels:

Fact 4.7 *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 act of suggestion, we can prove:

Fact 4.8 *The operation* $\mathfrak{M}_{\sharp\varphi}$ *preserves reflexivity and transitivity.*

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

On the other hand, applying our upgrades $\sharp\varphi$ can lead to loss of *connectedness* of the betternness order. Our earlier example in Section 4.1 already showed this: see Fig. 4.3. To us, this is no problem, since we are working with pre-orders. Indeed,

we see the fragility of connectedness under quite reasonable preference changes as an argument against adopting connectedness in general.

4.3.2 Upgrade and Model Transformation

Here is one more technical point about our mechanism. To a logician, the standard epistemic update $!\varphi$ of Chapter 2 essentially *relativizes* a model $\mathfrak{M}$ to a definable sub-model $\mathfrak{M}_{!\varphi}$ (see Definition 2.7 again). The relation between evaluation on both sides is expressed in the following standard result:

Fact 4.9 (relativization lemma) *Formulas φ hold in the relativized model iff their syntactically relativized versions were true in the old model:*

$$\mathfrak{M}_{!\varphi} \models \psi \quad \textit{iff} \quad \mathfrak{M} \models (\psi)^{\varphi}.$$

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

A very similar idea applies to betterness upgrade $\sharp\varphi$. This time, the relevant semantic operation on models is *redefinition of base relations*.[5] In this light, the reduction axioms for dynamic betterness logic 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 betterness 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:

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

4.3.3 A Program Format for Relation Change

The preceding observation is the key to a more general logic of preference dynamics. Reference [42] is a first systematic study along this line. The above relatively modest ordering change leaves the set of worlds the same, but it removes any preferences the agent might have for $\neg\varphi$ over φ among these. Clearly, it is just one of many possible "betterness transformers". To achieve greater generality, we use program notation from *propositional dynamic logic* (*PDL*, [104]).

---

[5] Reference [29] already noted 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.

For instance, as we said, we can write "suggesting that φ" ($\sharp\varphi$) as:

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

where R is the given input relation, while the operations $?\varphi$ test whether the relevant proposition φ or related ones hold. In particular, the disjunct $(?\varphi; R; ?\varphi)$ means that we keep all old betterness links that run from φ-worlds to φ-worlds. This definition is equivalent in *PDL* to the following:

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

Again this says, but now more compactly, that we keep all old R-links, except for those that ran from φ-worlds to $\neg\varphi$-worlds.

But then, given the observed diversity of possible preference changes, from weaker to stronger, we want to consider more general relation-changing operations. For instance, if one wanted to add links, rather than just subtract them, the above format would still work. E.g., the relation-extending stipulation

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

with $\top$ as the "universal relation", makes every φ-world better than every $\neg\varphi$-world.

Natural stronger preference changes occur in *belief revision* based on plausibility orders of worlds that show many similarities with betterness order. In this area, one prominent relation transformer is $\Uparrow\varphi$ ("radical revision with φ": cf. the dynamic logic in [29]). In preference terms, it is like a *strong command* revising the betterness order, say, by incorporating the wish of some over-riding authority.

Definition 4.10 (radical command upgrade) Given any modal betterness model $(\mathfrak{M}, s)$ and formula φ, the radical command upgrade $(\mathfrak{M}_{\Uparrow\varphi}, s)$ is the model with relations defined as follows:

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

This reflects the intended meaning of this transformation: All φ-worlds become better than all $\neg\varphi$-worlds, whether or not they were better before – but within these two zones, the old ordering remains.[6]

While suggestions and radical commands are extremes on a spectrum, other options exist. Here is one more notion form belief revision theory. *Conservative upgrade* puts only the *best* φ-worlds on top, underneath, the old ordering remains.[7] Again, this makes sense, now as a weaker command changing betterness order: The new best worlds will satisfy φ, but we leave the agent more of her own original betterness order. The latter is not irrelevant, since, as we have seen in Chapter 11, it encodes her conditional responses to further information. Differences with radical commands will then show up in judgments of "conditional betterness".[8]

---

[6] It is instructive to see the difference with $\sharp\varphi(R)$ in the above *PDL*-style format.

[7] Conservative upgrade is a radical command to produce, not φ, but $Best(\varphi)$.

[8] Such judgments occur in the literature on conditional obligation: see [96].

The above notions suggest a general syntax for definable betterness changes. In fact, we can use programs starting from tests for formulas in the language, weak and strict basic order relations as well as the universal relation, while allowing arbitrary unions and sequential compositions:

$$\pi := ?\varphi \mid R \mid R^{<} \mid \top \mid ; \mid \cup$$

These are interpreted as the standard *PDL* program operations of *test* $?\varphi$, *sequential composition*; and *choice* $\cup$.

Actually, our examples of suggestions and radical command upgrade used only special *flat forms* in this program format, without sequential composition of base relations. That is, they were unions of finite "trace expressions" of the form

$$`?\varphi_1; \{R, \top\}; ?\varphi_2; \{R, \top\}; \ldots',$$

where $\{R, \top\}$ stands for either R or $\top$. These flat forms will return later in our treatment of reason-based preference change.

Many further relation transformers can be defined in a *PDL* format. Of interest to us here is mainly the point that all drive an automatic formulation of dynamic completeness theorems. Reference [42] proved a general result to this effect, noting that the *PDL*-format allows us to push the dynamic modality inductively through the program structure, providing a computation of the shape of the recursion axiom. In our present setting, we extend this to the language with strict betterness modalities.[9]

Theorem 4.11 *Consider any dynamic program* π *as defined above. There is a complete dynamic betterness logic for this operation, and its reduction axioms for the weak and strict modalities can be computed automatically.*

Proof Consider any formula $\langle R := \pi(R)\rangle\varphi$, where we have made the intended relation change for the current betterness order R explicit in the dynamic modality.

The inductive reduction axioms when φ is a propositional variable or a Boolean complex are as in standard dynamic-epistemic logic, as presented in Chapter 2. Now consider the two modalities: i.e., φ is either $\langle R\rangle\varphi$ or $\langle R^{<}\rangle\varphi$. It suffices to show how these are affected, since the rest of the formula can be dealt with recursively, as displayed in the following equivalence:

$$\langle R := \pi(R)\rangle\langle R\rangle\varphi \;\leftrightarrow\; \langle \pi(R)\rangle\langle R := \pi(R)\rangle\varphi$$

Next, consider the inductive construction of the program π. We first rewrite any formula $\langle R := \pi(R)\rangle\varphi$ as far as possible by means of the well-known *PDL*-axioms for test, union and composition.

The remaining dynamic modalities for base programs involve either the identity $R := R$, and these modalities can evidently be dropped, or the atomic re-assignment $R := R^{<}$. We are done if we can show what happens in the latter case with both modalities that can follow, weak and strict.

[9] We do not consider dynamic actions that directly transform the strict betterness relation, since these are less intuitive, while also presenting some technical difficulties.

Here the case of $\langle R := R^{<}\rangle\langle R\rangle\varphi$ is immediate: We replace the prefixed weak modality $\langle R\rangle$ by the strict modality $\langle R^{<}\rangle$ that we have in our language anyway. Similarly, the other case of $\langle R := R^{<}\rangle\langle R^{<}\rangle\varphi$ appears to require a repeated strictness modality $\langle(R^{<})^{<}\rangle\varphi$. But this reduces to what we have:

$(R^{<})^{<} = R^{<}$.

Here is the simple Boolean calculation:

$$(R^{<})^{<}xy \leftrightarrow R^{<}xy \wedge \neg R^{<}yx \leftrightarrow R^{<}xy \wedge \neg(Ryx \wedge \neg Rxy)$$
$$\leftrightarrow R^{<}xy \wedge (\neg Ryx \vee Rxy) \leftrightarrow (R^{<}xy \wedge \neg Ryx) \vee (R^{<}xy \wedge Rxy)$$
$$\leftrightarrow R^{<}xy \vee R^{<}xy \leftrightarrow R^{<}xy$$ □

This concludes our general treatment of betterness transformers.

4.4 A Different Illustration: Default Reasoning

We have presented an upgrade mechanism for incoming triggers that changes betterness relations. We now illustrate how this method also works in quite different settings, viz. default reasoning.

4.4.1 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. An influential dynamic treatment of this sort of reasoning is [192], which makes the semantics of a default conditional 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, an instruction of the linguistic form "*Normally,* φ".

Now, in the perspective of this chapter, the same effect can be reached without changing the standard semantics of the underlying language, but by adding an explicit mechanism on top of that for getting the force of an act of default *assertion*. 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 procedure that will validate the preceding intuition. It is just the *radical command upgrade* defined and described above, but now reinterpreted as a plausibility change.[10]

[10] This "lexicographic" policy for belief revision was first suggested in [145].

Fact 4.12 *Relational default processing can be axiomatized completely.*

Proof By the method of Section 4.2, the key reduction axiom follows automatically from the given *PDL*-form, yielding

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

□

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.

But 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 or commands. We will provide a proposal to deal with this issue in the context of deontics in Chapter 11. One more 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. [154] for a good exposition, and we will explore the application of this idea in preference logics in Chapter 5.

Finally, 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. [91, 163], and Chapter 5). Indeed, the paper [29] 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.

4.4.2 Complication: Coherence and Conflicting Suggestions

Despite the technical analogies between information update and betterness upgrade, there are also intuitive differences. We end by discussing one of these.

Consider the intuitive notion of "coherence" in agency. In pure public announcement logics, the only relevant aspects of coherence for a sequence of assertions seem to be these:

(1) Do not make *inconsistent* and false assertions at the actual world; and, do not waste anyone's time.
(2) Do not make assertions which are *common knowledge* in the whole group, and which do not change the model.

But in combination with betterness 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 that which sequences of upgrades are coherent, in that they preserve the property of connectedness.

But in line with the main theme of this chapter, reality may be more *dynamic* than mere coherence maintenance. One often dynamically *resolves* conflicts in suggestions. One powerful way of doing this is 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 *DEL*, whether for update or upgrade. For the case of betterness upgrade, we will consider a possible solution in Chapter 11.

4.5 Conclusion

To conclude this chapter, we say a few words on what we have done, and what we have not. We have concentrated on the dynamics of betterness relations in changing possible worlds models. Our pilot example was a dynamic action of suggestion that upgrades betterness relations. We also saw how such a logic will automatically describe how lifted generic preferences transform under preference change. We then saw how a complete dynamic betterness logic results, driven by *DEL*-style reduction axioms. Next, we showed how this simple model can be extended to many other general relation-transforming operators, deriving reduction axioms automatically, provided the model changes are definable in a suitable program format. Finally, we showed how this same circle of ideas applies to areas like belief revision via change in plausibility relations, and to reasoning with defaults.

Nevertheless, some issues of importance have been left out here. First, in real agency, the notion of preference is often *entangled* with beliefs and knowledge. The delicate relation between these notions calls for further models and special attention. Likewise, for dynamics, we now have separate systems for preference and belief, but we also need to see how they interact. These phenomena have received little attention in dynamic logic circles so far, and we will give them a try in Chapter 5.

Secondly, we have represented the notion of preference qualitatively in terms of a binary betterness relation over possible worlds. But an equally dominant representation in the literature are quantitative utility functions, as in decision theory, game theory, and social choice theory. In such settings, preference change means numerical "valuation change". This quantitative approach is not the main line of this book, but there is at least the challenge of seeing whether the proposals that we have made here can survive the transition to such a richer setting. This will be the topic of Chapter 6, a sort of intermezzo in the main stream of this book.

Chapter 5
Entanglement of Preference, Knowledge and Belief

5.1 Introduction

Many of the meaningful scenarios where preference occurs have a mixed character. For instance, in making rational decisions, an agent does not just take her preferences into account, but also what she knows or believes about the outcomes of her actions. And in rational social decisions, she will also factor in what she knows about the preferences of others.[1] Knowledge, belief and preference are often deeply entangled in our reasoning.

One locus where this entanglement plays very clearly are "generic preferences" as lifts from the betterness relation, which naturally invite epistemic attitudes to involve. Recall that in Chapter 3, we defined a pure preference between propositions in the following manner:

$$\psi \trianglelefteq^{\forall\exists} \varphi := U(\psi \to \langle \leq \rangle \varphi). \qquad (Ubett)$$

A universal modality U was used in this definition. It suggests that all possible worlds in the model are involved in the comparison, as shown in Fig. 5.1.

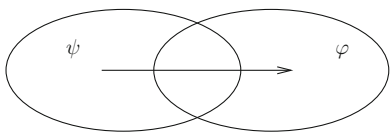


Fig. 5.1 Genetic preference defined by U and betterness

Basically, this is a comparison between all ψ-worlds and φ-worlds in the model. However, in a more realistic setting, the following restriction is natural: We only consider those possible worlds that are within our epistemic range. Thus, we only compare worlds that are epistemically accessible:

> For any ψ-world that is *epistemically accessible* to agent a in the model, there exists a world which is as good as that world, where φ is true.

[1] This phenomenon is called "interdependent preference" in the economic literature.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_5, © Springer Science+Business Media B.V. 2011

This can be pictured in the following manner Fig. 5.2:

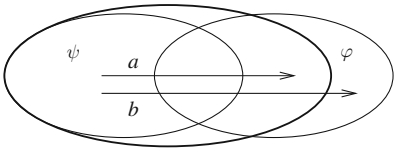


Fig. 5.2 Generic preference defined by K and betterness

The part inside the black circle stands for the epistemically accessible worlds. We see that only some of the φ-worlds are epistemically accessible. Note that the betterness relation has two possible cases: It has a-arrows which mean that the better φ-world is itself in the accessible part of the model, while b-arrows mean that the better φ-world need *not* be in the accessible part of the model.

We write the above explanation into a formal definition:

$$\psi \trianglelefteq^{\forall\exists} \varphi := K(\psi \to \langle\leq\rangle\varphi). \qquad (Kbett)$$

Comparing the definitions $(Kbett)$ and $(Ubett)$, we have simply replaced U with K. In fact, looking back at Chapter 3, this is a straightforward step to take, since we can simply combine knowledge with betterness operators, as we are indeed going to do in the next section.

This is just a first step in investigating possible ways of entangling preference, knowledge and beliefs. This chapter investigates these issues in more depth, along the following lines. Section 5.2 studies a simple way of combining knowledge and preference languages, by juxtaposition of epistemic and preference logic. We will explore how this combination affects the dynamics, and we will propose a new update mechanism in terms of "link-cutting". In addition, in Section 5.3 we will briefly discuss how these methods can be applied to the case of preference mixed with beliefs. Section 5.4 then explores a more intimate way of combining preference and beliefs, and a new merged model and logic will be proposed. In that logic, we will be able to talk about preference and plausibility relations at the same time. We end this chapter with some discussion of further forms of entanglement, and then conclude.

5.2 Juxtaposing Knowledge and Betterness

5.2.1 Static Logic

A direct way of combining preference and knowledge is to simply put epistemic logic and modal betterness logic together. The language is then as follows:

Definition 5.1 (epistemic betterness language) Take a set of propositional variables Φ, with p ranging over Φ. The epistemic betterness language is given by the following inductive syntax rule:

$$\varphi := \bot \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K\varphi \mid [\leq]\varphi \mid [<]\varphi \mid U\varphi.$$

Similarly, the new models combine epistemic and betterness models.

Definition 5.2 (epistemic betterness model) An epistemic betterness model is a tuple $\mathfrak{M} = (S, \sim, \leq, V)$, with S a set of possible worlds, $\sim$ the usual equivalence relation of epistemic accessibility, $\leq$ is a reflexive and transitive betterness relation over the worlds, and V a valuation for proposition letters.

Given our choice of epistemic betterness models, our epistemic betterness logic can be axiomatized completely in a standard modal style. It is essentially a **S5** system for knowledge plus a **S4** system for betterness (cf. [50]).

Theorem 5.3 *Epistemic betterness logic is completely axiomatizable w.r.t epistemic betterness models.*

Proof The proof is entirely by standard techniques from the modal literature. □

5.2.1.1 Expressive Power

A combined language like this can define many notions that make intuitive sense, such as having knowledge, or being ignorant, of someone's preferences, with combinations $K\langle\leq\rangle\varphi$. It can also express the opposite combination $\langle\leq\rangle K\neg\varphi$, the logical form of preferring to know it is not the case, say, that one has some unpleasant disease.

In particular, there is the issue of epistemically "loading" the very notion of preference itself. As we have seen earlier in Chapter 3, the pure modal betterness part of this language, with the help of the universal modality, can express a variety of natural notions of preference between propositions (generic preference), including the original one proposed by von Wright, as well as other natural options qua quantifier combinations. But with our additional epistemic operators, we can also express the interplay of betterness and knowledge. The following examples represent (i) an intuition of self-reflection of "preference", and (ii) an unfortunate but ubiquitous phenomenon:

- $\langle\leq\rangle\varphi \rightarrow K\langle\leq\rangle\varphi$: Positive Betterness Introspection.
- $\langle\leq\rangle\varphi \wedge K\neg\varphi$: Regret.

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

Fact 5.4 *An epistemic betterness frame $\mathfrak{F} = (S, \sim, \leq)$ makes the Betterness Introspection Axiom $\langle\leq\rangle\varphi \rightarrow K\langle\leq\rangle\varphi$ true iff it satisfies the following condition:*

$$\forall s \forall t \forall u : (s \leq t \wedge s \sim u \rightarrow u \leq t).$$

In other words, we can add this introspection axiom to the axiomatic system and get a slightly richer logic. However, in this chapter, we will keep it only as an option. As we will show in the next subsection, the above frame condition cannot be preserved in what we consider a natural betterness dynamics. We will return to mixed epistemic-preference principles again soon.

5.2.2 Dynamic Logic and Some New Operations

Given the static epistemic betterness logic, knowledge update and preference upgrade do not lead wholly separate lives in our setting. For instance, if we want to model the realistic phenomenon of "regret" about worlds that are no longer epistemic 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.[2]

One way of doing this is by redefining the update of Chapter 2 for public announcement as a milder relation-changing operation of "link-cutting". This time, instead of the earlier public announcement $!\varphi$, we write the relevant update action as

$$\dagger\varphi$$

and we write the updated model as $\mathfrak{M}_{\dagger\varphi}$ in order to distinguish it from what we obtained by eliminating worlds. The correct semantic operation for $\dagger\varphi$ is this:

Definition 5.5 (link-cutting update) The link-cutting public update model $\mathfrak{M}_{\dagger\varphi}$ is the original model $\mathfrak{M}$ with its worlds and valuation unchanged, but with accessibility relations $\sim$ replaced by the following version without any crossing between the φ- and $\neg\varphi$-zones of $\mathfrak{M}$:

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

Link-cutting has some interesting features. For instance, link cutting in the current model is the same for the announcements $\dagger\varphi$ and $\dagger(\neg\varphi)$: Both remove links between φ-worlds and $\neg\varphi$-ones. This is reflected in the following truth condition of the link-cutting action, as well as in the valid principles of the dynamic epistemic betterness logic below.

Definition 5.6 Given an epistemic model $\mathfrak{M}$, the truth definition for the link-cutting action modality is the following:

$$(\mathfrak{M}, s) \models [\dagger\varphi]\psi \quad \text{iff} \quad \mathfrak{M}_{\dagger\varphi}, s \models \psi.$$

Theorem 5.7 *The following formulas are valid principles of combined dynamic epistemic betterness logic in its link-cutting version:*

(1) $\langle\dagger\varphi\rangle p \leftrightarrow p$.
(2) $\langle\dagger\varphi\rangle\neg\psi \leftrightarrow \neg\langle\dagger\varphi\rangle\psi$.
(3) $\langle\dagger\varphi\rangle(\psi \wedge \chi) \leftrightarrow (\langle\dagger\varphi\rangle\psi \wedge \langle\dagger\varphi\rangle\chi)$.
(4) $\langle\dagger\varphi\rangle\langle K\rangle\psi \leftrightarrow (\varphi \wedge \langle K\rangle(\varphi \wedge \langle\varphi!\rangle\psi) \vee (\neg\varphi \wedge \langle K\rangle(\neg\varphi \wedge \langle\varphi!\rangle\psi)$
(5) $\langle\dagger\varphi\rangle\langle\leq\rangle\psi \leftrightarrow \langle\leq\rangle\langle\dagger\varphi\rangle\psi$.
(6) $\langle\dagger\varphi\rangle E\psi \leftrightarrow E(\langle\dagger\varphi\rangle\psi \vee \langle\neg\dagger\varphi\rangle\psi)$.
(7) $\langle\sharp\varphi\rangle p \leftrightarrow p$.
(8) $\langle\sharp\varphi\rangle\neg\psi \leftrightarrow \neg\langle\sharp\varphi\rangle\psi$.

---

[2] More concretely, in the games of Chapter 12, it is essential that agents be able to reason counterfactually about their preferences, even in states of the game of which they know that these will not occur.

(9) $\langle\sharp\varphi\rangle(\psi \wedge \chi) \leftrightarrow (\langle\sharp\varphi\rangle\psi \wedge \langle\sharp\varphi\rangle\chi)$.
(10) $\langle\sharp\varphi\rangle\langle K\rangle\psi \leftrightarrow \langle K\rangle\langle\sharp\varphi\rangle\psi$.
(11) $\langle\sharp\varphi\rangle\langle\leq\rangle\psi \leftrightarrow (\neg\varphi \wedge \langle\leq\rangle\langle\sharp\varphi\rangle\psi) \vee (\langle\leq\rangle(\varphi \wedge \langle\sharp\varphi\rangle\psi))$.
(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 $\langle\dagger\varphi\rangle$ and $\langle\leq\rangle$, expresses the fact that epistemic update does not change any betterness relations. As we saw, the usual updates $!\varphi$ and the link-cutting updates $\dagger\varphi$ 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 in the above is different, as $E\varphi$ can still refer to worlds after the update which used to be $\neg\varphi$.

Next in the list comes a similar set of reduction principles for the upgrade operation, but we have seen these earlier already in Chapter 4. □

The link-cutting update has a number of advantages. It was first proposed, in [175] and [130] for modeling 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

$$Kp \wedge \langle\leq\rangle\neg p.$$

Here is an intuitive reading: "Yes, I know that p, but it would be better if it weren't..." The modified link-cutting update allows us to have this consistently. More generally speaking, in the setting of preference update, we cannot eliminate possible worlds– and instead, what we do is to change the preference relations between them. Link-cutting typically fits this spirit.

5.2.2.1 Loss of Positive Introspection

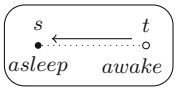
A combined epistemic betterness logic contains natural mixed epistemic-betterness principles. One obvious candidate is *betterness introspection*: Agents know their preferences, at least in the sense of knowing what is better for them. Now the issue becomes whether such principles survive natural update operations. We find interesting phenomena.

We have shown in Chapter 4 that the operation $\sharp\varphi$ preserves reflexivity and transitivity. But as we have stated concerning introspection, unfortunately, our upgrade cannot preserve this property, witness the following scenario:

Example 5.8 (upgrade complications) Consider Fig. 5.3 below. There are two worlds "asleep" and "awake". We do not know if we are sleeping or awake. Initially, we prefer being asleep, and we know our preference.

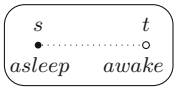
Now an upgrade happens, suggesting that real waking life is not so bad after all. Then we still do not know if we are sleeping or awake, but at the "awake" world we

Fig. 5.3 Before suggestion



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" (Fig. 5.4). Betterness Introspection fails!

Fig. 5.4 After suggestion



In some settings, betterness introspection seems plausible, and a desirable property of models to be preserved. We can then change the 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 betterness introspection. The latter would then be the "sensible" coherent series of suggestions.

This concludes our brief exploration of betterness merged with knowledge. Things get more realistic, and also more interesting, when we merge betterness and *belief*. Our next section explores this entanglement.

5.3 Connecting Belief with Betterness

Before turning to the entanglement of belief and betterness, let us first talk briefly about beliefs, whose semantics shows many analogies with betterness models for preference. The following discussion draws together a few points that have occurred in earlier chapters.

5.3.1 Belief Statics and Dynamics on Their Own

Recall a semantics that we have discussed briefly already in Chapter 3, when exploring the expressive power of our modal base language. Our betterness pre-orders are formally exactly the same as *plausibility models* for a language of belief. And indeed, as mentioned briefly in Chapter 3, they immediately support a doxastic language with an absolute belief operator B and more generally, conditional beliefs $B^{\psi}\varphi$. The key truth conditions are as follows:

$\mathfrak{M}, s \models B\varphi$ iff $\mathfrak{M}, t \models \varphi$ for all worlds t which are minimal for the ordering $\lambda xy.\ \leq_s xy$.

$\mathfrak{M}, s \models B^{\psi}\varphi$ iff $\mathfrak{M}, t \models \varphi$ for all worlds t which are minimal for $\lambda xy.\ \leq_s xy$ in the set $\{u \mid \mathfrak{M}, u \models \psi\}$.

Here the truth condition for the absolute operator B is essentially the same as in **KD45**-models, with the "accessible worlds" of the latter system being the most plausible ones. But we can compare less plausible worlds, too – and this is crucial to understanding the logic of conditional belief. The logic of the latter is essentially the *minimal conditional logic* over pre-orders: see [83] for an complete axiomatization, along the lines of earlier work by Burgess and Veltman.

A few points are worth noticing here. The first is the difference in our basic languages for what are formally the same models. For the betterness models of Chapter 3, we took a simple base modality, while here, we take more complex minimality versions with a more complex pattern of quantification pattern in their truth condition. The latter seems crucial, being tied up with the well-known non-monotonicity of conditional beliefs in their antecedent.

Still, the analogy is actually quite fruitful. In particular, it makes sense to introduce an ordinary universal modality $[\preceq]\varphi$ on plausibility models after all, read as "in all worlds that are at least as plausible as the current one". If we take the set of all worlds plausibility related to the current world (less, equally, or more plausible) as the latter's epistemic range, then this universal modality is intermediate in force between knowledge and belief. It has been called *safe belief* in the doxastic literature, and [18] show how it makes a lot of sense as a notion of belief that is stable under receiving truthful information.[3] Indeed, as we have seen technically in Chapter 3, for many purposes, safe belief suffices for *defining* absolute and conditional belief in the minimality sense. For instance, conditional belief $B^{\psi}\varphi$ became the following complex modal formula

$$B^{\psi}\varphi := U(\psi \rightarrow \langle\preceq\rangle(\psi \wedge [\preceq](\psi \rightarrow \varphi))).$$

This way the logic of conditional belief lies encoded under translation in our modal base logic.

Finally, in many recent papers, the semantic setting is a bit more complex, combining epistemic and doxastic features. In that case, plausibility relations live inside the epistemic range of the current world, entangling modalities of knowledge and belief. In particular, one often finds an assumption of epistemic introspection for beliefs, corresponding to an assumption that the plausibility order is the same for any two epistemically indistinguishable worlds. We will not pursue this combination here, but it does seem the natural setting eventually.

Finally, with all these analogies, the *dynamics of belief change* can be studied using our techniques for preference. This was done in [29], which 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

[3] Safe belief has roots going back to the computational tradition on agency, cf. [174] and to philosophical logic, cf. [179]. Interestingly, [15] cited our paper [42] on preference dynamics as one of their inspirations.

$\Uparrow A$. For instance, van Benthem's reduction axiom for beliefs after radical revision is:

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

Adding a reduction axiom for conditional belief from [29] to the valid principles for plausibility upgrade, we immediately get a complete dynamic logic.

We end with a somewhat sobering observation. While these axioms for belief change under hard or soft information look mysterious, and may be hard to discover, they actually follow quite simply from our earlier analysis in Chapter 3 and 4!

Fact 5.9 *The correct reduction axioms for absolute and conditional belief are automatically computable from the axioms given for preference change in Chapter 4.*

Proof The reason is a simple combination of our observations so far. First, conditional belief is definable in terms of basic safe belief. Next, radical upgrade is definable as a *PDL* program transformer on models. Therefore, by Theorem 4.11, we can automatically compute its matching reduction axiom for the base modality. Finally, given the definition of conditional belief in terms of base modalities, we can use the latter reduction axiom to derive a reduction axiom for conditional belief – in the same style as we did for lifted preferences in Chapter 4. It is easy to check that this analysis derives, and to some extent explains, van Benthem's original dynamic logic of belief change. □

5.3.2 Beliefs Together with Betterness

Having explored the many analogies between belief and preference, we can combine these two formally congenial notions. Beliefs combine with betterness in the same way as we had for knowledge and betterness.

Now let us plunge in immediately, and resume our earlier discussion of entangled preference, this time based on belief rather than knowledge. First, the comparison can be expressed in the following:

> 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.

Formally this gives us a new notion of generic preference:

$$\psi \trianglelefteq^{\forall\exists} \varphi := B(\psi \rightarrow \langle\leq\rangle\varphi). \qquad (Bbett)$$

Figure 5.5 illustrates what we have in mind. In this picture, worlds lie ordered according to their plausibility, as in Lewis' spheres for conditional logic. The inside of the black circle depicts the most plausible worlds. We consider the ψ-worlds in this area, and again distinguish two sorts of betterness: Relations of "type a" stay inside the most plausible region, relations of "type b" go outwards to the less plausible, or even highly implausible regions.

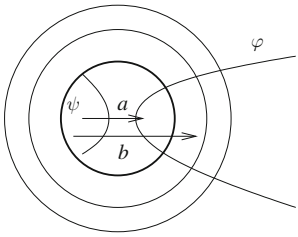


Fig. 5.5 Generic preference defined by B and betterness

For the record, we now define the obvious combined models as follows:

Definition 5.10 (doxastic betterness model) A doxastic betterness model is a tuple $\mathfrak{M} = (S, \preceq, \leq, V)$, with S a set of possible worlds, $\preceq$ a doxastic relation of "at least as plausible as", and $\leq$ our earlier relation of "at least as good as", with V again a valuation for proposition letters.[4]

Definition 5.11 (doxastic betterness language) Take a set of propositional variables Φ, with p ranging over Φ. The doxastic betterness language is given by the following inductive syntax rule:

$$\varphi := \bot \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid B\varphi \mid B^{-}\varphi \mid B^{\psi}\varphi \mid [\leq]\varphi \mid [<]\varphi \mid U\varphi.$$

Note that we have both strict and non-strict versions in the language for belief and betterness operators, we will need the strict version soon when we compare our model with an earlier proposal in [123]. The truth condition for the strict belief operator is defined as follows:

$\mathfrak{M}, s \models B^{-}\varphi$ iff for all worlds t if $s \prec t$, then $\mathfrak{M}, t \models \varphi$.[5]

Regarding the axiomatization, it is a combination of **KD45** system for beliefs and **S4** system for betterness.

---

[4] References [16] and [18] 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.

[5] 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

$$B^{\psi}\langle\leq\rangle\varphi. \qquad (Bbett')$$

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.

5.3.2.1 Comparison with an Earlier Proposal

A similar model with two relations was also proposed in [55]. It works like this: $\mathfrak{M} = (S, \leq_P, \leq_N, V)$, where S is a set of possible worlds, V a valuation function and $\leq_P$, $\leq_N$ are two transitive connected relations. $x \leq_P y$ means "y is as preferable as x" and $x \leq_N y$ means "y is as normal as x". [55] defines *conditional ideal goal* (IG) by combining preference and normality relations:

Definition 5.12 $\mathfrak{M} \models IG(\varphi \mid \psi)$ iff $Max(\leq_P, Max(\leq_N, Mod(\psi))) \subseteq Mod(\varphi)$

Intuitively, this says that φ is an ideal goal with condition ψ if and only if the best of the most normal ψ worlds satisfy φ.

Fact 5.13 *The notion of an ideal goal in [55] is definable in the modal doxastic betterness language.*

Proof Here is how we can define ideal goals in our language:

$$IG(\varphi \mid \psi) := (\psi \wedge \neg\langle B^- \rangle \psi) \wedge \neg\langle < \rangle (\psi \wedge \neg\langle B^- \rangle \psi) \rightarrow \varphi \qquad \square$$

Next, following [55, 123] defined preference using the above kind of semantic model, and studied its properties in more detail. Here is their definition:

Definition 5.14 $\mathfrak{M} \models \psi < \varphi$ iff for all $w' \in Max(\leq_N, Mod(\psi))$ there exists $w \in Max(\leq_N, Mod(\varphi))$ such that $w' <_P w$.

Readers will see the similarity between this definition and our definition $(Bbett)$. Indeed, we can easily obtain the following fact:

Fact 5.15 *The notion of entangled preference in [123] is definable in our doxastic betterness language. Moreover, all properties identified in that paper are formally derivable in our complete doxastic betterness logic.*

Proof To prove the first part, it is sufficient to see the following:

- The $\forall\exists$ format of the definition is the same.
- "the most normal φ-worlds" are defined by "$\varphi \wedge \neg\langle B^- \rangle \varphi$" in our language.

We omit the mechanics of the formal proofs for the second claim. $\square$

This concludes our discussion of preference combined with belief. The match seems natural, and it even extends beyond what we said here. For instance, merely combining what we have already seen in Chapter 3 and 4, we can now describe two kinds of *preference dynamics*. One comes about through changes in the plausibility relation, triggered by hard or soft information, the other comes about through suggestion – or command-style changes in the betterness relations. The *DEL* methodology applies to both cases, and the result is a very rich theory of information and evaluation change.

5.4 Deeper Entanglement: Merged Belief and Betterness

5.4.1 Static Logic

Still, there is something left to be desired. The entangled definitions ($Kbett$) and ($Bbett$) share a common feature, namely, an arrow (of type b) leading to a better φ-world can go *outside of the accessible or most plausible part* of the model. 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.

Nevertheless, equally intuitive to us, sometimes we do want to just look at better alternatives inside the relevant epistemic or doxastic zone. We have just seen such considerations in [123]. Likewise, [39] discuss the "normality sense" of ceteris paribus preference, restricting preference relations to just the normal worlds for the agents. In this section, we therefore make a further proposal, making the entanglement of preference and belief even closer by intersecting the plausibility and betterness relations directly in the model.

Definition 5.16 (merged doxastic betterness model) A *merged doxastic betterness model* is a tuple $\mathfrak{M} = (S, \leq, \preceq, \leq \cap \preceq, V)$, with S a non-empty set of possible worlds with doxastic and betterness relations, but also the relation $\leq \cap \preceq$ as the intersection of the relations "at least as good as" and "at least as plausible as", with V again a valuation for propositional variables.

The original language had separate modal operators B and $[\leq]$, but now we extend it with a new modality H. The new language is defined as follows:

Definition 5.17 (merged doxastic betterness language) Take a set of propositional variables Φ and a set of nominals Nom, with p ranging over Φ and $i \in Nom$. The *merged doxastic betterness language* is given by the following rule:

$$\varphi := \bot \mid p \mid i \mid \neg\varphi \mid \varphi \wedge \psi \mid B\varphi \mid [\leq]\varphi \mid H\varphi \mid U\varphi.$$

One possible reading of $H\varphi$ is that "Hopefully, it is the case that φ". Note that we have also introduced nominals from hybrid logics, for technical convenience in what follows. Let us call the above language $\mathcal{L}$. In addition, following [84] and [79], we define one more notion, for technical reasons.

Definition 5.18 (necessity form in $\mathcal{L}$) Let \$ be a symbol not belonging to $\mathcal{L}$. We define inductively the notions of necessity forms in $\mathcal{L}$.

$$l := \$ \mid \psi \rightarrow l \mid Bl \mid [\leq]l \mid Ul \mid Hl.$$

Definition 5.19 (truth conditions) Given a merged doxastic betterness model $\mathfrak{M}$, and a world $s \in S$, we define $\mathfrak{M}, s \models \varphi$ (formula φ is true in $\mathfrak{M}$ at s) by induction on the construction of φ. We omit the standard cases:

$$\mathfrak{M}, s \models i \quad \text{iff} \quad V(i) = \{s\}.$$

$$\mathfrak{M}, s \models H\varphi \quad \text{iff} \quad \text{for all } t \text{ with both } s \leq t \text{ and } s \preceq t, \text{ it holds that } \mathfrak{M}, t \models \varphi.$$

With this new language over these enriched models, we can define one more natural notion of generic preference:

$$\psi \trianglelefteq^{\forall\exists} \varphi := B(\psi \rightarrow \langle H\rangle\varphi). \qquad (BH)^{6}$$

Intuitively, 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 betterness relations restricted to the plausible part of the model. In terms of Fig. 5.5, only arrows of "type a" remain.

Now let us quickly look at the expressive power of the modal language with the new "hopefully" operator H. The following example shows that it is more expressive than the language with separate operators B and $[\leq]$. Can the notion of preference in (BH) be defined in the original language with (iterated) modal operators B and $[\leq]$ only? As we know from general modal logic, intersection modalities are not invariant under bisimulation (cf. [50]). Indeed, here too, the answer is negative:

Fact 5.20 *$B(\psi \rightarrow \langle H\rangle\varphi)$ $(*)$ is not definable in the standard bimodal language with modal operators B and $[\leq]$.*

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 depicted in Fig. 5.6.

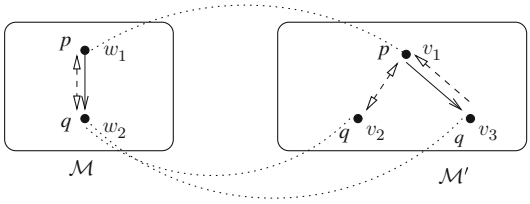


Fig. 5.6 Bisimilar models

The betterness relation $\leq$ is pictured by solid lines with arrows, and the plausibility relation $\preceq$ 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.

Also, we have $\mathfrak{M}, w_1 \models B(p \rightarrow \langle H\rangle q)$, 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 $\mathfrak{M}, w_1 \models \varphi$, since $\varphi \leftrightarrow (*)$. Because $\mathfrak{M}$ and $\mathfrak{M}'$ are bisimilar, we would then have

---

[6] Actually, this same move would apply to Definition $(Kbett)$ as well. Requiring that the better worlds stay inside the accessible worlds, we would have:

$$\psi \trianglelefteq^{\forall\exists} \varphi := K(\psi \rightarrow \langle \sim \cap \leq\rangle\varphi) \qquad (Kbett').$$

This means that we keep only the a-arrows in Fig. 5.2.

$\mathfrak{M}', v_1 \models \varphi$. So we should also have $\mathfrak{M}', v_1 \models B(p \to \langle H\rangle q)$. But instead, we have $\mathfrak{M}', v_1 \not\models B(p \to \langle H\rangle q)$, 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. Here is an axiomatization for merged doxastic betterness models:

Theorem 5.21 *The merged doxastic betterness logic is completely axiomatized by the usual propositional tautologies,* ***S5****-principles for U,* ***KD45****-principles for B,* ***S4****-principles for* $[\leq]$ *and the following principles and inference rules:*

(1) $H(\varphi \to \psi) \to (H\varphi \to H\psi)$
(2) $H\varphi \to HH\varphi$
(3) $H(H\varphi \to \varphi)$
(4) $E(i \wedge \varphi) \to U(i \to \varphi)$
(5) Ei
(6) $U\varphi \to B\varphi$
(7) $U\varphi \to [\leq]\varphi$
(8) $U\varphi \to H\varphi$
(9) $B\varphi \to H\varphi$
(10) $[\leq]\varphi \to H\varphi$
(11) $(\langle B\rangle i \wedge \langle\leq\rangle i \to \langle H\rangle i) \wedge (\langle H\rangle i \to \langle B\rangle i \wedge \langle\leq\rangle i)$
(12) *Modus ponens: If* $\vdash \varphi$ *and* $\vdash \varphi \to \psi$, *then* $\vdash \psi$.
(13) *Generalization for the operator B, U,* $[\leq]$ *and H*
(14) *If* $\vdash l(\neg i)(i \in Nom)$, *then* $\vdash l(\bot)$, *where l is in necessity form.*

Proof The proof can be found in the recent note [204]. □

Here ends our study of static logics of informationally entangled preferences.

5.4.2 Dynamic Logic

But now we have a challenge. For the first time, we have significantly extended our modal base language. Will the dynamic methodology of this book survive this generalization?

Let us now move to dynamic events of preference change, and consider changes to the merged relations. To simplify things, we only look at our three characteristic actions: radical revision $\Uparrow\varphi$ that changes plausibility relations, suggestion $\sharp\varphi$ that changes betterness relations, and standard public announcement $!\varphi$ that changes the domain of worlds. As it happens, *DEL* reduction axioms still work:

Theorem 5.22 *The following equivalences are valid:*

(1) $\langle\sharp\varphi\rangle\langle H\rangle\psi \leftrightarrow (\varphi \wedge \langle H\rangle(\varphi \wedge \langle\sharp\varphi\rangle\psi)) \vee (\neg\varphi \wedge \langle H\rangle\langle\sharp\varphi\rangle\psi)$.
(2) $\langle\Uparrow\varphi\rangle\langle H\rangle\psi \leftrightarrow (\varphi \wedge \langle H\rangle(\varphi \wedge \langle\Uparrow\varphi\rangle\psi)) \vee (\neg\varphi \wedge \langle H\rangle(\neg\varphi \wedge \langle\Uparrow\varphi\rangle\psi)) \vee (\neg\varphi \wedge \langle\leq\rangle(\varphi \wedge \langle\Uparrow\varphi\rangle\psi))$.
(3) $\langle!\varphi\rangle\langle H\rangle\psi \leftrightarrow \varphi \wedge \langle H\rangle\langle!\varphi\rangle\psi$.

Proof We only explain the most interesting Axiom 2 as an illustration. Assume that $\langle \Uparrow \varphi \rangle \langle H \rangle \varphi$. Recall that radical revision $\langle \Uparrow \varphi \rangle$ only changes the plausibility relation, leaving the betterness relation intact. The new plausibility relation can be written as follows, as we have seen in Chapter 4:

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

Seen from the initial model, we can therefore distinguish three cases, and these are just the three disjuncts displayed on the right-hand side. Note that for the last one we only need to insert the old betterness relation $\langle \leq \rangle$, since the plausibility relation $(?\neg\varphi; \top; ?\varphi)$ is new. □

In particular, what we see here is that intersection modalities, while asking for some special devices like nominals in the static logic, do not pose any new difficulties in terms of dynamic reduction axioms.[7]

By adding the above axioms to the complete static logic introduced in the previous section, we obtain a complete dynamic logic for entangled preference changes. Thus, the *DEL*-methodology of this book also works in this extended setting.

5.5 Discussion and Conclusion

Generic preferences revisited In this chapter, we have proposed several new definitions for the notion of generic preference. They involve either knowledge plus betterness, or belief plus betterness, or new merges like "hope". Dynamic changes to plausibility relation and betterness relation can still be dealt with in *DEL* approach, yielding reduction axioms for the new entangled operators. Following our calculation of generic preferences in the pure modal betterness language, calculating reduction axioms for the new mixed cases is a routine exercise.

Other forms of entanglement Of course, there is a long tradition of combining preference and beliefs in decision theory ([113, 166]). There most models rely on a numerical representation where utility and uncertainty are commensurate. For instance, an agent may not know the outcomes of her actions, but will use a probability distribution over outcomes to compute an *expected value* of an action. The latter notion, explained in any textbook on probability and utility, deeply entangles the agents' betterness relations and her beliefs about possible outcomes. By contrast, our logical approach is qualitative, though the reader will see a more quantitative version of our ideas in *DEL* style in Chapter 6 below. We leave a thorough comparison of our work with probabilistic systems for another occasion.[8]

---

[7] To be fully precise, we would also have to show how the various operations considered here deal with the *nominals* extending our static language. We refer the reader to [204] for technical details of this particular topic.

[8] Dynamic epistemic methods are certainly compatible with probabilistic approaches. See [17, 37, 165] for some recent probabilistic versions of *DEL*.

Conclusion This chapter has investigated the issue of entanglement, i.e., various ways of combining preference with epistemic knowledge or doxastic beliefs. The first model proposed was a simple juxtaposition of betterness logic and epistemic logic. We studied both its static and its dynamic logic. Link-cutting, as a less drastic mechanism of information update for this new setting, was extensively discussed. Then we applied our methods to the case of betterness and beliefs, showing first how belief statics and dynamics can be fruitfully developed in analogy with our preference logics. Then we showed how a similar approach worked for belief-entangled preference. Next, we moved one step further, proposing a new entangled preference modality intersecting betterness and plausibility at the level of the semantic models. The expressive power of this new language was discussed, and then we showed that both its static and dynamic logic still yield to the general techniques of this book.

Overall, this chapter has shown how the methods of this book can develop quite rich systems of information and evaluation dynamics working together. After a quantitative intermezzo on these same themes in the next chapter, we will return to preference itself in Part IV, suggesting that we need, and can get, richer styles of modeling than we have given so far.

Chapter 6
Intermezzo: A Quantitative Approach

6.1 Introduction

The notion of preference arises from comparisons of alternatives. To formalize such comparisons and study them, there are two ways to go. We can represent preference in terms of qualitative binary relations, as we did in our modeling in Chapter 3, following a long logical tradition ([83, 100, 122] and [92]). Or we can introduce a utility or evaluation function which will assign values to the alternatives being compared. The latter quantitative method has been dominant in many research areas, e.g., game theory and social choice theory (cf. [74, 113, 114] and [63]).

Of course, there is a philosophical background behind the utility approach. The doctrine of utilitarianism saw the maximization of utility as a moral criterion for the organization of society.[1] In economics, utility is a measure of relative satisfaction, and in this sense, preference is then considered in the context of consequentialism.

The distinction in approaches makes no difference for the importance of our key topics so far, such as preference change. Given a measure of utility, one may speak meaningfully of increasing or decreasing utility, and indeed, explain economic behavior dynamically in terms of attempts to increase one's utility.

In this chapter, we want to briefly explore how our concerns fare in a quantitative setting. Our treatment will be brief and anecdotal, since our main aim is showing how the qualitative methods of this book also work in principle for quantitative settings. But there are a few things to be said at the outset, where we assume that our readers already know some basics.

First, it is not our intention to explain everything about utility theory and evaluation functions. We will just use ideas and results from utility theory to better understand preference and preference change in the context of information processing. Secondly, the connection between preference and utility to keep in mind is this:

[1] This position was held by famous utilitarians such as Jeremy Bentham (1748–1832) and John Stuart Mill (1806–1876). It has even been claimed that this position is found millennia earlier in China with Mozi.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_6,

An alternative with a higher utility value is preferred to one with a lower value.[2] Thirdly, assuming ordinal utilities, the magnitude of utilities has no meaning, only their ranking. Still, we often want to model strength of preferences in many real situations: The numbers matter, and we will explore this perspective, too.

Our eventual aim is to show how our dynamic epistemic methodology fits with quantitative approaches. We will introduce ideas of evaluation into the *DEL* framework and see how to model preference change then, making a junction with utility theory.

Here is a simple quantitative scenario extending an earlier example:

Example 6.1 (buying a house, revisited) Suppose that Alice plans to buy a new apartment. There are two candidate apartments d_1 and d_2 available, located in different places. She has her own preference judgement based on her current knowledge. To mark her evaluation difference, she assigns two numbers to d_1 and d_2, respectively. A newspaper article that "the government is planning to build a park near d_1" may *increase* her value for d_1. In contrast, getting to know that the crime rate is going up in the neighborhood of d_1 may *decrease* her value for d_1.

The idea is that one starts off with initial values for the options, and keeps *scoring* in accordance with the new information, adding points if the information has a positive influence on the option, or dropping points in case it has a negative effect. A number zero would be added when there is no effect, say, when the new information is irrelevant. But of course, point scores could also just happen without informational triggers, like the intrinsic betterness changes that we have studied earlier. Both kinds of evaluation change can then trigger preference changes.

The chapter is organized as follows. An epistemic evaluation logic will be introduced in Section 6.2, with an axiomatization for which we prove completeness. In Section 6.3 we will study the dynamics of preference in this framework, using a new mechanisms of numerical product update. A matching dynamic epistemic evaluation logic will be presented, too. With this in place, we then discuss how to parametrize the update mechanism to various types of agents, while still getting reduction axioms. Section 6.4 is about the expressiveness of the evaluation language, while the issue of preference strength is considered as well. Section 6.5 will look at a few examples in a deontic context with a quantitative view.[3] We end the chapter with the conclusions that we draw from all this.

6.2 Epistemic Evaluation Logic

Following [176], a language of graded belief modalities has been introduced to indicate the strength or degree of beliefs in [10]. Possible worlds in the model

[2] A well-known result is that a preference relation that is complete, reflexive, transitive and continuous can be represented by a continuous utility function. In this chapter, we only consider representable preferences.

[3] We will discuss deontic applications in much greater detail in Chapter 11.

were then ordered according to their plausibility, which is represented by numerical devices. Here we take a similar idea for preference, and define a simpler language with numerical evaluation atoms developed in [130], which is more workable and perspicuous.

Definition 6.2 (epistemic evaluation language) Let a finite set of proposition variables Φ and a set of agents N be given. The epistemic evaluation language $\mathcal{L}_{\mathcal{E}}$ is defined by the rule:

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid q_a^m \mid K_a\varphi \quad \text{where } p \in \Phi, a \in N, \text{and } m \in \mathbb{Z}.$$

Note that the language is an extension of epistemic language. A propositional constant q_a^m is added to the language for each agent $a \in N$ and each value $m \in \mathbb{Z}$. The intended reading of the formula q_a^m is "the agent a assigns the state where she stands the value at most m". In what follows, again we will try to suppress the subscription when it is clear in context.

Definition 6.3 (epistemic evaluation models) An evaluation model for the epistemic evaluation language is a tuple $\mathfrak{M} = (S, \sim, v, V)$ such that S is a non-empty set of states, $\sim$ is an epistemic equivalence relation on S, v is an evaluation function assigning each state an element from $\{-\infty\} \cup \mathbb{Z} \cup \{\infty\}$[4], and V is a function assigning to each propositional variable p in Φ a subset $V(p)$ of S.

Semantically speaking, each possible world is associated with some number, denoting the value that the agent has towards that world. As we explained in Section 6.1, evaluation functions induce a total ordering in an obvious way, namely, from $v(s) \leq v(t)$ we can obtain $s \leq t$. In this manner, we are making use of the information about the qualitative ordering encoded in the evaluation functions. However, we will see that the quantitative information part will play a big role in many situations in the following sections. For instance, considering information about the intensity of preference will lead to a new definition of bisimulation.

Definition 6.4 (truth conditions) Suppose s is a state in a model $\mathfrak{M} = (S, \sim, v, V)$. We can inductively define the notion of a formula φ being true in $\mathfrak{M}$ at state s. Here we omit those standard clauses and only give one for the propositional constant:

$\mathfrak{M}, s \vDash q^m$ iff $v(s) \leq m$, where $m \in \mathbb{Z}$.

For the sake of comparison, we give the definition for $B^m\varphi$ in the language $\mathcal{L}_A$ of [10] as follows:

$\mathfrak{M}, s \vDash B^m\varphi$ iff for all $t \in S$: if $s \sim t$ and $v(t) \leq m$, then $\mathfrak{M}, t \vDash \varphi$.[5]

---

[4] In [10] the range is natural numbers up to a maximal element (Max). The values are normalized to Max. For me the distance between the numbers seems essential, so normalization is not an option. Similarly I like to be able to subtract unrestrictedly.

[5] Let us look at the relation between $\mathcal{L}_A$ and $\mathcal{L}_{\mathcal{E}}$. From $\mathcal{L}_A$ to $\mathcal{L}_{\mathcal{E}}$, we can define a translation: a formula of the form $B^m\varphi$ is translated into $K(q^m \to \varphi)$. This is to say that in the language $\mathcal{L}_{\mathcal{E}}$, we can

Epistemic Evaluation Logic (*EEL*) consists of the following axioms and derivation rules:

(1) All propositional tautologies.
(2) $K(\varphi \to \psi) \to (K\varphi \to K\psi)$.
(3) $K\varphi \to \varphi$.
(4) $K\varphi \to KK\varphi$.
(5) $\neg K\varphi \to K\neg K\varphi$.
(6) $q^m \to q^n$ for all $m \leq n \in \mathbb{Z}$.
(7) If $\vdash \varphi$ and $\vdash \varphi \to \psi$, then $\vdash \psi$.
(8) If $\vdash \varphi$, then $\vdash K\varphi$.

Theorem 6.5 (soundness) *EEL is sound over epistemic evaluation models.*

Axiom 6 is essential here. According to the natural properties that $\mathbb{Z}$ has, we can easily conclude that the preference is reflexive, complete (total) and transitive.

We assume the standard notion of a formal proof. In case a formula φ is provable in *EEL*, we write $\vdash_{EEL} \varphi$.

Theorem 6.6 (completeness) *EEL is complete with respect to evaluation models.*

Proof The proof is standard. First we define the canonical model as follows: $\mathfrak{M}^c = (S^c, \sim, v, V)$ is the structure with

- $S^c = \{s_S\colon S \text{ maximal } EEL\text{-consistent set}\}$.
- $\sim = \{(s_S, s_T)\colon S/K \subseteq T\}$. where $S/K = \{\varphi\colon K\varphi \in S\}$.
- $v(s_S) = \min\{m : q^m \in S\}$ (∞ if $\{m : q^m \in S\}$ is empty, $-\infty$ if $\{m : q^m \in S\} = \mathbb{Z}$).
- $s_S \in V(p)$ iff $p \in S$.

We need to show that

$$\varphi \in T \text{ iff } \mathfrak{M}^c, s_T \models \varphi.$$

This can be done by induction on the structure of the formula φ. We only consider the case of the constant q^m:

($\Rightarrow$) Assume $q^m \in T$. We have $v(s_T) \leq m$. Then by Definition 6.4, we get $M^c, s_T \models q^m$.

($\Leftarrow$) Assume $M^c, s_T \models q^m$. We know $q^{v(s_T)} \in T$ and $v(s_T) \leq m$. By axiom 6, $q^{v(s_T)} \to q^m$. So, we get $q^m \in T$. This is to say that we have proved that

Every *EEL*-consistent set Γ of formulas is satisfied in some epistemic model.

---

express the same notions as [10] without introducing additional belief operators. This advantage leads to the much simpler completeness proof we will see. It becomes even more prominent when constructing reduction axioms for dynamics in the later sections. On the other hand, we can easily translate $\mathcal{L}_{\mathcal{E}}$ back into $\mathcal{L}_A$: q^m will be $\neg B^m \bot$, which means that $\mathcal{L}_A$ and $\mathcal{L}_{\mathcal{E}}$ are equivalent.

Completeness of epistemic evaluation logic now follows in the usual manner.[6] □

Having set up the base language for evaluation models, we now proceed to the dynamic superstructure that we have in mind.

6.3 Dynamic Epistemic Evaluation Logic

6.3.1 Evaluation Product Update

Before defining things formally, let us consider the example in our Section 6.1 first.

Example 6.7 Assume that in the initial model S_0, agent a has the same value for s and t where d_1 would be chosen at s and d_2 at t. This means there is no particular preference for Alice, she gives 0 to both of them, pictured below (Fig. 6.1):

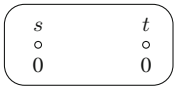


Fig. 6.1 Initial model S_0

Afterward, the newspaper tells Alice that "the government is planning to build a park near d_1" (p). This positively affects the value of s in the model S_0, but unfortunately has no effect on t. The initial model S_0 is then updated to S_1 (Fig. 6.2):

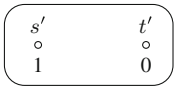


Fig. 6.2 Updated model S_1

In the new model S_1, clearly, a would prefer d_1 over d_2, as the value for s' is now greater than that for t'.

The story goes on, the new information "the crime rate is going up in the neighborhood of d_1"(q) causes values to decrease. The resulting model is depicted in Fig. 6.3.

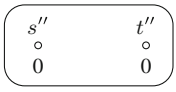


Fig. 6.3 Updated model S_2

With these value changes, preference changes accordingly: As before, agent a has no preference between d_1 and d_2.

[6] The main reason why this argument can remain so simple, compared to earlier numerical systems, is our use of propositional constants for values of worlds.

This example shows how incoming information changes values of the states constantly. To model a phenomenon of this sort, we define a new sort of "event models" in the *DEL* style (cf. [14, 65]), that describe the change to take place.

Definition 6.8 (evaluation event model) A evaluation event model is a tuple $\mathcal{E} = (E, \sim, v, PRE)$ such that E is a non-empty set of events, $\sim$ is a binary epistemic relation on E, v is an evaluation function assigning each action an element from $\mathbb{Z}$, PRE is a function from E to the set of all epistemic propositions.

Based on the values they assign to events, the evaluation functions v indicate how agents (are to) evaluate events. Note that this is a major change as compared with standard uses of evaluation: We do not just evaluate static states of affairs, but we can also evaluate actions or events![7]

Definition 6.9 (evaluation product update) Let an evaluation model $\mathfrak{M} = (S, \sim, v, V)$ and an evaluation event model $\mathcal{E} = (E, \sim, v, PRE)$ be given. The evaluation product update model is then defined to be the model $\mathfrak{M} \otimes \mathcal{E} = (S \otimes E, \sim', v', V')$ such that

- $S \otimes E = \{(s, e) \in S \times E\}$.
- (s, e) $\sim'$ (t, f) iff both $s \sim t$ and $e \sim f$.
- $v'(s, e) = v(s) + v(e)$ (Addition rule).
- $V'(p) = \{(s, e) \in S \otimes E : s \in V(p)\}$.

Note that we keep all world/event pairs (s, e) represented in the new model, as even the non-realized options may be ones that we still have regrets about. For the evaluation update clause, we have simply taken the *sum* of the values for the previous state and for the event.[8] This "Addition Rule" is best understood by looking at Example 6.7 again, though the evaluation event model there is quite simple. In general, an evaluation event model can be much more complex, and also, the update process can go on for a very long time. What remains constant at each stage is that agents prefer things with a higher score.

However, several issues remain to be discussed here. First of all, the *sources* of information may not all be equally reliable. In order to propose a realistic evaluation update rule, the *reliability* of information must be taken into account. A related issue concerns the different *forces* of information. In multi-agent settings, the same information may have a different force for different agents. For instance, agent a may take a piece of information seriously, while agent b does not do so. These two aspects are parameterized in the following new update rule.

Definition 6.10 (parametrized update rule) Let $\mu(e)$ be a reliability function, and $\lambda(e)$ a relative force function. The domains of these two functions are the set of

[7] Actually, there are two different plausible interpretations here. One either thinks of the event as a command to change one's evaluation of some current state, or as something which itself has a value that leads to a prescribed value change in the world where the event takes place.

[8] Other numerical rules are possible, but most of our later points would apply then as well.

events, and the ranges of these functions are $\mathbb{N}$.[9] Given the value for the previous state s and event e, the new value for state (s, e) is defined by the following equation:

$$v'(s, e) = v(s) + v(e) \cdot \mu(e) \cdot \lambda(e).$$

Returning to the first step of Example 6.7, suppose that agent a only half trusts what the newspaper said, namely $\mu(e) = 5$. Moreover, the relative force of the park building information is 4, i.e. $\lambda(e) = 4$, which means that she thinks it rather important. Then the value of s' in the model S_1 would be calculated as

$$v'(s, e) = 0 + 1 \cdot 5 \cdot 4 = 20$$

With the parametrized rule, we can understand better how information is being processed. But things need not stop here. One could propose other evaluation update rules to interpret more complex situations. For example, the agent may give more weight to the previous state (behave conservatively), which seems to call for a parameter associated with the value for s in the above rule, as was proposed for belief revision of diverse agents in [130] and [132]. Or in some situations, one needs to consider dependencies between information that comes later and that comes earlier (cf. [37]). We will not pursue these issues here.

6.3.2 Dynamic Epistemic Evaluation Logic

We are now ready to define a logic for dynamic evaluation update mechanisms. In this section we confine ourselves to the Addition Rule only.

Definition 6.11 (dynamic epistemic evaluation language) Let a set of propositional variables Φ, and a set of events E be given. The dynamic epistemic language is defined by the rule

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid q^m \mid K\varphi \mid [e]\varphi \quad \text{where } p \in \Phi, e \in E, \text{ and } m \in \mathbb{Z}.$$

We will not include the usual action operations like composition, choice, or iteration that form composite action scenarios. These do not add anything essentially related to our numerical setting. What *is* typical for the latter setting are formulas of the form $[e]q^m$, for which we will find reduction axioms as follows:

Theorem 6.12 (soundness) *Dynamic epistemic evaluation logic (DEEL) consists of the following formulas, and it is sound w.r.t. evaluation product update models:*

(1) $[e]p \leftrightarrow p$.
(2) $[e]\neg\varphi \leftrightarrow \neg[e]\varphi$.
(3) $[e](\varphi \wedge \psi) \leftrightarrow [e]\varphi \wedge [e]\psi$.
(4) $[e]K\varphi \leftrightarrow PRE(e) \rightarrow \bigwedge_{f \in E}\{K[f]\varphi : e \sim f\}$.

[9] In practice, one will normally choose a small natural number, say, between 0 and 10, to denote the reliability or the relative force.

(5) $[e]q^m \leftrightarrow q^{m-v(e)}$.

Proof To prove the validity of the above axioms, we consider two models: $(\mathfrak{M}, s)$ and $(\mathfrak{M} \otimes \mathcal{E}, s)$ before and after the update. Axiom (1) says that the update will not change the objective valuation of atomic propositions. And Axioms (2) and (3) are just Boolean operations, easy to see. Axiom (4) was explained in Chapter 2, what is different here is that the general event model was considered. In Axiom (5), the formula $[e]q^m$ says that, in $\mathfrak{M} \otimes \mathcal{E}$, the agent a assign the value at most m to the world s where she stands. According to the Addition rule, the value of s in $(\mathfrak{M} \otimes \mathcal{E}, s)$ is the sum of the value for s in $\mathfrak{M}$ and that for e in $\mathcal{E}$. Thus the right value for the world s in $\mathfrak{M}$ is at most $m - v(e)$. This is what Axiom (5) says. □

Theorem 6.13 (completeness) *The logic DEEL is in fact axiomatized completely by the above reduction axioms.*

Proof We have seen the soundness of the above reduction axioms. Note that they are all equivalences, which makes them clearly sufficient for eventually turning every formula from the dynamic language into a static one. Then we can use the completeness theorem for our static evaluation language in Section 6.2. □

One final issue remains to be discussed: Do other update rules define a complete logic, and in particular, the parametrized rule? There is no general results here. But the parametrized rule does suggest the following reduction axiom. Although it seems a bit clumsy, its validity can be proved in a similar way to Axiom (5).

$$[e]q^m \leftrightarrow PRE(e) \rightarrow q^{m-v(e)\cdot\mu(e)\cdot\lambda(e)}$$

However, once we go further, and introduce weights for the previous state, this job becomes harder. If the update rule is simply algebraically expressible, we can still get a complete logic, though clearly a simple subtraction will no longer work.

6.4 Excursion: Bisimulation for Evaluation Languages

Our numerical languages show many further analogies with the modal framework of this book. To get a good understanding of the expressiveness of the evaluation language presented in Section 6.2 we look at its behaviour under *bisimulation*, a fundamental notion in modal logics. First we formulate a standard notion of bisimulation for evaluation models[10]:

Definition 6.14 (evaluation bisimulation) Let $\mathfrak{M} = (S, v, V)$ and $\mathfrak{M}' = (S', v', V')$ be two evaluation models. A non-empty binary relation $Z \subseteq S \times S'$ is called an evaluation bisimulation between $\mathfrak{M}$ and $\mathfrak{M}'$ if the following conditions are satisfied:

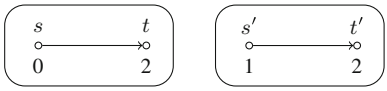
[10] The conditions for the epistemic relations $\sim$ are omitted, as they are routine.

(1) If sZs' then s and s' satisfy the same propositional variables.
(2) If sZs' and $v(s) \leq v(t)$ (or $s \preccurlyeq t$), then there exists t' in $\mathfrak{M}'$ such that tZt' and $v'(s') \leq v'(t')$ (or $s' \preccurlyeq t'$) (the forth condition).
(3) If sZs' and $v'(s') \leq v'(t')$ (or $s' \preccurlyeq t'$), then there exists t in $\mathfrak{M}$ such that tZt' and $v(s) \leq v(t)$ (or $s \preccurlyeq t$) (the back condition).

While this notion looks plausible, we do encounter a phenomenon unlike standard *DEL*. In a dynamic setting, this notion does not quite suffice!

Example 6.15 (numerically bisimilar models) From the viewpoint of the above evaluation bisimulation, it makes sense to identify the two models $\mathfrak{M}$ and $\mathfrak{M}'$ below, where we mark worlds by their values. After all, the pure preference pattern is the same in both. This would be in accordance with ordinal utility theory (Fig. 6.4).

Fig. 6.4 Models $\mathfrak{M}$ and $\mathfrak{M}'$



But our intuition shows that the values make a difference. Consider the event model $\mathcal{E}$ which updates all φ-worlds (s in the pictures) with 1 each time it is applied. Applying $\mathcal{E}$ once to the model on the left keeps the preference intact, but on the right, it voids it. All this seems to suggest that we need a new notion of bisimulation definition for evaluation models to express the *strength of preferences*. Our proposal uses the following notion:

Definition 6.16 (distance) The distance between two possible states s and t in an evaluation model is defined as $\mathcal{D}(s, t) = \mid v(s) - v(t) \mid$.

In Example 6.15 the distance between s and t is 2 in the model on the left, but it is 1 on the right. Now we can define a more sensitive notion:

Definition 6.17 (distance bisimulation) Let $\mathfrak{M} = (S, v, V)$ and $\mathfrak{M}' = (S', v', V')$ be two evaluation models. A non-empty binary relation $Z \subseteq S \times S'$ is called distance bisimulation between $\mathfrak{M}$ and $\mathfrak{M}'$ if the following conditions are satisfied:

(1) If sZs' then s and s' satisfy the same propositional variables.
(2) If sZs', $s \leq t (t \leq s)$ and $\mathcal{D}(s, t) = k$, then there exists t' in $\mathfrak{M}'$ such that tZt', $s' \leq t' (t' \leq s')$ and $\mathcal{D}(s', t') = k$ (the forth condition).
(3) If sZs', $s' \leq t' (t' \leq s')$ and $\mathcal{D}(s', t') = k$, then there exists t in $\mathfrak{M}$ such that tZt', $s \leq t (t \leq s)$ and $\mathcal{D}(s, t) = k$ (the back condition).

As usual, we say two evaluation models are *bisimilar* when there is some evaluation bisimilation linking two states in the two models. Intuitively, if the *same efforts* are made to get from one state to another in each model, then the two models are bisimilar.

This means that with the notion of comparative distance, we can say things like "d_1 is preferable over d_2 more than d_1 is preferable over d_3". This simply means $\mathcal{D}(s_1, s_2) > \mathcal{D}(s_1, s_3)$ in the model, where d_1, d_2 and d_3 are chosen in s_1, s_2 and

s_3, respectively. This metric quality is something that most languages of qualitative preference are not able to express.

We will not follow this line of investigation here, but it seems quite interesting to relate our system to the modal languages for geometry studied in [13], or to the numerical similarity relations in [194] and [99].

6.5 Excursion: Numerical Measures in Deontics

We now continue with a further use of our numerical systems. While we will study deontics at greater length in Chapter 11, it also appeals to us here for its concrete interpretation of the ideas introduced so far. Our aim is to show how, in this area, the logical issues in this chapter correspond to questions of independent interest.

Deontic logic: statics and dynamics Deontic logic (cf. [8]) started as the study of assertions of obligation like "it ought to be the case that φ" (written $O\varphi$) emanating from some moral authority. The underlying intuition of interpreting such sentences is the following: It ought to be the case that φ if φ is true in *all best possible worlds* as seen from the current one. This naturally suggests a deontic betterness ordering among worlds, linking deontics to preference. We will see in a moment how this motivates a richer quantitative version as well.

Likewise, we can think of the deontic setting dynamically: Obligations may change due to incoming new information, or they can change through moral commands, treated as actions themselves. There is a whole tradition along these lines, viewing deontic actions as programs of some sort, including [141, 142, 185, 203], and others.

In particular, the recent paper [200] takes the dynamic epistemic paradigm to obligation changes brought about by acts of commanding in a multi-agent context. Here is the key reduction axiom proposed in [199]:

$$[!_a\varphi]O_a\psi \leftrightarrow O_a(\varphi \rightarrow [!_a\varphi]\psi),$$

where the intended interpretation of $O_a\varphi$ is "it is obligatory for the agent a ($\in N$) that φ", and $[!_a\varphi]$ is intended to represent the action of commanding an agent a to see to it that φ.

Yamada's system can be translated into the qualitative relation-changing version of preference update presented in Chapter 4 (cf. [42]). But here, we are after an enrichment that makes a lot of sense in a setting of deontic actions.

Force of commands and parametrized numerical update Commands often come with a weight, and these weights may differ for various reasons. Either the strength of the moral authority is greater or larger, but also: Maybe the authority herself has a range of imperative strength, from mild suggestions to full-blown orders. Here is where our earlier evaluation event models apply. We can now indicate the "weight" of a command in terms of numerical points, as pictured in the following event model (Fig. 6.5):

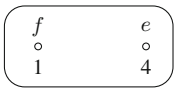


Fig. 6.5 Commands with weight

where command e comes with more strength than f does.

The earlier update mechanism for evaluation update now applies to obligation change as well, but with a more refined view of the effects. In particular, it brings new insights into a difficulty that has often been discussed: what to do with *conflicting commands*. Let us look at a variation of the key example in [199]:

Example 6.18 (Conflicts among authorities) You are reading an article in the office that you share with your two bosses and a few other colleagues. It is a hot summer noon, the temperature is above 30 degree Celsius. You can open the window, turn on the air conditioner, or you can concentrate on your reading and ignore the heat. Then your boss A commands you to open the window, but boss B immediately commands you not to do that. What effects do their commands have on the current situation? Which one do you obey?

Reference [199] handles this problem through a theorem of the form

$$[!_a(\varphi \wedge \neg\varphi)]O_a\psi \qquad (\text{“}DeadEnd\text{”})$$

It says that contradictory commands lead to an obligatory dead end, an impasse in the agent's situation. But this implicitly rules one important aspect of the situation, i.e. *the hierarchy of authorities*. Your two bosses may well stand at different levels of authority, and you may refuse to open the window if boss B is in a higher position than A. This shows that in a deontic setting, managing conflict is much more than managing consistency. To model the possible contradictory commands carried by different authorities, our current system provides at least one new way-out. It is through the following rephrased update rule:

Definition 6.19 (parametrized deontic update rule) Consider any *DEL*-style event model. Let $\eta(e)$ be an “authority function”, and $\lambda(e)$ a “relative force function”. The domains of both these two functions are the set of events in the model, and the ranges are the natural numbers $\mathbb{N}$. Given the value for the previous state s and the new event e, the new value for state (s, e) is defined as follows:

$$v'_a(s, e) = v_a(s) + v_a(e) \cdot \eta(e) \cdot \lambda(e).^{11}$$

By introducing a hierarchy of authorities into the above update rule, we actually deal with the problem of relative authority and conflicting commands *within* the logic. One promising way to take this issue further would be an explicit hierarchy of sources, and this is exactly what we will do in the priority-based analysis of preference in Chapters 7 through 9.[12]

[11] This also makes immediate sense in a multi-agent context, employing relative forces. One agent a may take the boss's commands seriously, whereas agent b may not.

[12] Another relevant approach are the *DEL* models for *trust* developed in [109].

Numerical default reasoning In line with these deontic examples, numerical dynamics also makes sense for a topic that we already discussed in a qualitative setting in Chapter 4, viz. *default reasoning*. Here, too, agents receive new information which does not necessarily eliminate worlds, but changes their evaluations of these. A typical example is the instruction "Normally, φ" in [192], which changes the preference ordering between worlds so as to give the φ worlds a higher position. While we will revisit the qualitative aspects of this connection at great length toward the end of Chapter 11, here we just make a point about a numerical dimension.

To model default-style evaluation update with a "normally" statement, we might take an event model $\mathcal{E}$ including two events "see φ", "see $\neg\varphi$" with different values, say 1 and 0. Executing the product update with $\mathcal{E}$ then leads to a new model where the φ-worlds have each gained one point, upgrading their position in the agent's pattern of plausibilities.

This rule is different from our earlier upgrade rules in that it does not make all φ-worlds automatically better than all $\neg\varphi$-worlds. It all depends on their previous scores.[13] In this way, the earlier dynamic evaluation language becomes a sort of refined numerical default language, where the expression ["*see* φ"] ψ plays the role of a default conditional "if φ then ψ". A complete evaluation default logic can be deduced directly from our general logic *DEEL*. But *DEEL* seems to be much richer than standard default logics, since by varying event values in $\mathcal{E}$, one can describe the behavior of a whole *family* of different default conditionals. It all depends on which numerical strengths an agent wishes to assign to the antecedents.[14]

6.6 Conclusion

We have presented a quantitative semantics for preference in terms of evaluation models. A new language with propositional value-constants was proposed, which turned out to be both concise and expressive. Moreover, we saw how this quantitative perspective suggests a more refined way of dealing with preference changes when processing new information. For this purpose, we generalized the standard dynamic-epistemic mechanism of product update, proposing a new Addition Rule, later also in a parametrized version, to model subtleties of value change. A matching complete dynamic epistemic evaluation logic was presented. Next, we looked at forms of bisimulation for the evaluation language, and found how these can be used to express strength of preferences. We then applied these ideas in deontic settings, proposing a solution to the well-known problem of contradictory obligations through conflicting commands.

---

[13] This is closer to Veltman's intuition that worlds are ordered by how many stated regularities they have satisfied in a longer discourse.

[14] As for a comparison with the dynamic semantics for defaults in [192], I suspect that the latter can be embedded into *DEEL*, but not vice versa.

This chapter is just a first step towards a full-scale quantitative study of preference and preference change. But we hope that, even at this stage, the reader will have seen how the qualitative paradigm of this book can be taken further than she may have thought.

This chapter also ends our studies on modal betterness models of preference. From the next chapter onward, we will focus on *reasons* for preference, leading to richer priority-based structures and their static and dynamic logics.

Part IV
Preference from Priorities

Chapter 7
Preference from Priorities: Static Logic

In this chapter, we will change the atmosphere of our topics and our logical methods a bit: Worlds will make place for objects, modal logic for first-order logic, and there will be differences in style as well. Eventually, however, all will fit back into one uniform paradigm for this book.

7.1 Introduction

In his pioneering book discussed in Chapter 1, von Wright distinguished three types of preference according to the objects that we are comparing: (i) (the use of) one instrument is preferred to (the use of) another instrument, (ii) one way of doing a thing is preferred to another way of doing the same thing, or (iii) one state of affairs is preferred to another state of affairs ([197] p. 12). Let us keep this distinction for the moment, though this is not exhaustive, and different types of preference can often be translated to each other.

If we consider the preference models used in Part III as mainly concerning states of affairs, this chapter will add one more type: preference between objects. Comparing objects naturally leads us to think of properties those objects have, and the reasons for our preferences. Thus, conceptually, the approach taken in this part is different from that of the previous one. Let us immediately single out its distinctive characteristics. Most previous work has taken preference to be a primitive notion of betterness between worlds, 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 reasonably derived from a more basic source, which we will call a *priority base*. In this manner we get two levels: the priority base, and the preference among objects derived from it. This richer 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 a priority base: A good overview can be found in [62]. In what follows, we will adopt one inspired by linguistic optimality theory.

When preference is employed to compare alternatives, one requirement we impose is that we consider only mutually exclusive alternatives. Objects are, of

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_7,

course, congenitally mutually exclusive. Although the priority base approach is particularly well suited to compare preference between objects, it can be applied to the comparison of other types of alternatives as well. In the next chapter, we will show how to apply the priority base approach to propositions.

When comparing objects, the kind of situation to be thought of is again our running example of buying a house:

Example 7.1 (buying a house) 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 criteria 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 fact whether the alternatives have the desirable properties, and also by the given order of importance for the properties.

In other words, Alice's preference regarding houses is derived from the priority order of the properties that she considers. This chapter aims at developing a logic to model such situations.

There are several points to be stressed before we start, in order to avoid misunderstandings. First, our intuition of a priority base is related to earlier ideas on graded semantics, in particular, the spheres semantics of [127] for conditionals. We take a more syntactical approach in this chapter, but that is largely a question of taste.[1] Secondly, we will mostly consider a linearly ordered priority base. This is simple, giving us a quasi-linear order of preference among objects. But our approach can be adapted to the more general partially ordered case, as we will indicate at the end of the chapter. Indeed, eventually, that is the general framework we will advocate, for instance, in Chapter 10. Finally, although the two-level perspective may look unfamiliar, it 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 results that show that any model of these modal logics can be seen as having been constructed from a priority base. These *representation theorems* provide a bridge to the usual purely modal completeness results.

The chapter is structured as follows. In Section 7.2, we start with a simple language to study the rigid case in which the priorities lead to a clear and unambiguous preference ordering. In addition, we show that the priority sequence approach is equivalent to Lewis' sphere semantics. In Section 7.3 we review some basics about ordering. Next, in Section 7.4 a complete preference logic is proposed and a proof is presented of a representation theorem for the simple language. In Section 7.5, we discuss how one can generalize our approach from linear orders to partially ordered priority bases, and what comes last are a few conclusions.

[1] In Chapter 10, we will discuss cases where the syntactic view is really richer.

7.2 From Priorities to Preference

7.2.1 Priority-Based 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 predicates, simple formulas, usually quantifier free or even variable free.

Given a priority base, there are various ways to get preference. The definition below is directly inspired by optimality theory. 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.[2] To take a neutral stance we use the words priority sequence instead of constraint sequence.

Definition 7.2 (priority sequence) 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.

---

[2] 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.

Definition 7.3 (preference) Given a priority sequence of length n, two objects x and y, $Pref(x, y)$ is defined as follows:

$Pref_1(x, y) := C_1(x) \wedge \neg C_1(y)$,
$Pref_{k+1}(x, y) := Pref_k(x, y) \vee (Eq_k(x, y) \wedge C_{k+1}(x) \wedge \neg C_{k+1}(y)), k < n$,
$Pref(x, y) := Pref_n(x, y)$,[3]

where the auxiliary binary predicate $Eq_k(x, y)$ stands for $(C_1(x) \leftrightarrow C_1(y)) \wedge \cdots \wedge (C_k(x) \leftrightarrow C_k(y))$.[4] In the following we will write $\underline{Pref}$ for the non-strict version of preference.

In Example 7.1, 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., $Pref(d_1, d_2)$.

The method introduced in the above easily applies when the priorities become graded. Take the Example 7.1, 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. This is very natural in real life. The original priority sequence $C(x) \gg Q(x) \gg N(x)$ may change into

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

Preference derived from such a priority sequence gets refined, we would have a better grasp on the situation we are in.

7.2.2 Syntactic Versus Semantic Views

As we mentioned at the beginning, we have chosen a syntactic approach expressing priorities by formulas. If we switch to a semantic 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 7.1 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 ([127] pp. 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

[3] Unlike in Chapter 8 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*.

[4] This way of deriving an ordering from a priority sequence is the "leximin ordering" of [62].

are linearly ordered by inclusion. To compare with the priority base we switch to a syntactical variant of sphere semantics, a sequence of formulas $G_1, \ldots, G_m$ 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.

Theorem 7.4 *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 in the following:

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

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 Fig. 7.1. □

Remark 7.5 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.

Remark 7.6 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

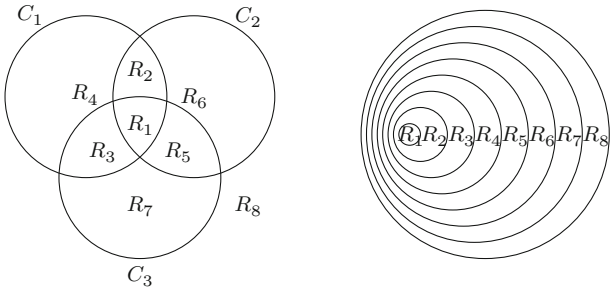


Fig. 7.1 C_i, R_i, and G_i

one of these ways, called *best-out ordering* in [62], 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 to Example 7.1, taking this option means that we only get the conclusions $\underline{Pref}(d_1, d_2)$ and $\underline{Pref}(d_2, d_1)$: d_1 and d_2 are equally preferable, since after observing that $\neg Q(d_1)$, $\neg \underline{Q}(d_2)$, Alice won't consider the factor N at all.

7.3 Order: Some Basics

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 translate 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 $\leq$ 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 $<$ 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 $<$ as an irreflexive, transitive, asymmetric relation. Then $<$ is definable in terms of $\leq$ by use of (2), but not $\leq$ in terms of $<$. That is clear from the picture below (Fig. 7.2), an irreflexive, transitive, asymmetric relation cannot distinguish between the two given orderings.

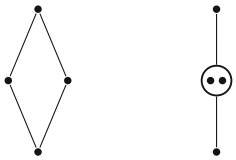


Fig. 7.2 Incomparability and indifference

One needs an additional equivalence relation $x \sim y$ to express that x and y are elements in the same cluster and they are indifferent. We will write Eq for $\sim$, expressing two elements are equivalent. $x \sim y$ 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 7.5. 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 use (6) to obtain a non-strict quasi-order from a strict one and $\sim$. However, we will see in the next chapter 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 $<$ as well as $\sim$ imply $\leq$.

7.4 Preference Logic and a Representation Theorem

Clearly, no matter what the priorities are, the non-strict preference relation has the following general properties:

(1) $\underline{Pref}(x, x)$,
(2) $\underline{Pref}(x, y) \vee \underline{Pref}(y, x)$,
(3) $\underline{Pref}(x, y) \wedge \underline{Pref}(y, z) \rightarrow \underline{Pref}(x, z)$.

(1), (2) and (3) express reflexivity, connectedness and transitivity, respectively. Thus, $\underline{Pref}$ is a quasi-linear relation; it lacks antisymmetry.

Unsurprisingly, (1), (2) and (3) 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:

Definition 7.7 (reduced preference language) Let Φ be a set of propositional variables, and D be a finite domain of objects, the reduced preference language is defined as follows:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \underline{Pref}(d_i, d_j) \quad \text{where } p \in \Phi \text{ and } d_i \in D.$$

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

(1) $\underline{Pref}(d_i, d_i)$,
(2) $\underline{Pref}(d_i, d_j) \vee \underline{Pref}(d_j, d_i)$,
(3) $\underline{Pref}(d_i, d_j) \wedge \underline{Pref}(d_j, d_k) \rightarrow \underline{Pref}(d_i, d_k)$.

We call this axiom system **P**.

Theorem 7.8 (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.[5] □

Remark 7.9 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 $\mathfrak{M}$ 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 $\mathfrak{M} \models \varphi_{ij}(d_i)$and $\mathfrak{M} \models \neg\varphi_{ij}(d_j)$. 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.

[5] 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.

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.

It is good to point out here that if one considers the objects as worlds and replaces the monadic predicates by propositional variables, the results so far can be restated in hybrid logic (see e.g., [49]), provided formulas are restricted to be quantifier free. This has advantages and disadvantages. Stating the results in hybrid logic has the advantage that it makes the results more directly comparable to those of other papers that consider preference between worlds or propositions rather than objects. The approach here allows more complex predicate-logical formulas to be used as constraints. Also, it allows generalizations to belief contexts in the following chapter. At this time we do not see how to obtain these benefits in hybrid logic.

7.5 Discussion and Conclusion

We now indicate how the approach taken here can be taken further.

Generalizing to partially ordered priorities A new situation occurs when there are several priorities of incomparable strength. Take Example 7.1 again, but this time, instead of considering three properties, Alice also takes "transportation convenience" into account. For her, though, neighborhood and transportation convenience are incomparable. Abstractly speaking, this means that the priority sequence is now *partially ordered*. This is the more general case of priority-based preference, when criteria are incomparable, or in conflict. We will see many illustrations in later chapters.

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.

Definition 7.10 (preference from partial-ordered priorities) We define $Pref_n(x, y)$ by induction, where $\{n_1, ..., n_k\}$ is the set of immediate predecessors of n.

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

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 better-known lexicographic orders from the literature: cf. [89] and [6].[6] We will come back to the partially ordered priorities in great detail in Chapter 10, where we will shift the presentation to a mathematically more elegant general format of *priority graphs* instead of priority sequences.

Much more can be said about the logical theory of partially ordered priorities and their induced betterness orders, but we leave this to later in this book.

Priorities versus generic preferences Finally, we ask a question that may have occurred to the reader already. How are priority sequences of propositions related to the orderings of propositions produced by *set lifting* of betterness relations, as in Chapter 3? 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. Here are some simple results of this sort on the earlier $\trianglelefteq^{\forall\exists}$-lifting. Priority order and lifted generic preference can coincide then when we work with special sets of worlds:

Definition 7.11 A set X is upward closed if

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

Fact 7.12 *Consider only sets X that are upward closed. We define*

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

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 when transformations are repeated. Given a priority order and its induced betterness order, when can the former be retrieved as a quantifier lift of the latter? And also, given a world order, and some lift, say $\trianglelefteq^{\forall\exists}$, used as a priority order on the powerset $\mathcal{P}(S)$, when can the relation on S be retrieved as the derived order of $\trianglelefteq^{\forall\exists}$? Answering such 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.

Conclusion In this chapter we put the *reasons* for preference at center stage, and we studied how to then represent preference and its logic. Our main intuitions were about comparing objects, though everything we have said also applies to preference

[6] More discussion on the relation between partially ordered priorities and G-spheres, is found in [128] and, when it is unordered, [121].

between worlds. We showed how preference over objects can be derived from underlying priorities, ordered sequences of properties of these objects. Using a fragment of the first-order language, we proposed a set of valid principles for preference, and we matched the usual completeness result with a representation theorem telling us which object orderings are the ones induced by linear priority sequences. Finally, we discussed some further issues, in particular, extensions of our preference definitions to cases where the priorities are partially ordered (a topic that will be pursued in much greater detail in Chapter 10), and connections between ordered priorities and the generic preferences of Chapter 3, showing that in some particular situations they can coincide indeed.

All our results in this chapter presuppose that an agent has complete information when forming her preference. A more realistic situation is that agents may be uncertain about certain things, relying on her beliefs to make preference judgements. This is the topic we will take up in the next chapter.

Chapter 8
Belief-Based Preference

8.1 Introduction

In this chapter, we resume the issue of "entanglement" of preference and information-based notions like knowledge and belief, that we have studied already in Chapter 5. How does this play when preference comes with richer priority structure?

To plunge right in, let us consider a variation of Example 7.1:

Example 8.1 (buying a house under uncertainty) 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. Consider two houses d_1 and d_2 that Alice hopes to choose from. Alice only has partial information. Let us assume that she *believes* that d_1 and d_2 have the following properties: $C(d_1)$, $C(d_2)$, $\neg Q(d_1)$, $\neg Q(d_2)$, $N(d_1)$ and $\neg N(d_2)$.

The definition of preference proposed in Chapter 7 does not apply here anymore, as beliefs have entered now. Alice's decision is not determined by her complete information, but by her beliefs under uncertainties. In a more general sense, this allows us to consider more complex scenarios. For instance, do we believe certain properties from the priority base to apply or not? Or even more dramatically, can we form a priority base on the basis of our beliefs? Handling uncertainties of this kind calls for a combination of a doxastic language and a preference language.

For this purpose, the preference language defined in the previous chapter will be extended now with belief operators $B\varphi$. When we do this, 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 interpreting modal predicate logics (cf. [80]). What we are using in this context is just a very limited part of such a language.[1] We will take the standard modal system **KD45** as the logic for belief, though we are

[1] It would be interesting to consider what more a full doxastic predicate logic language can bring to our preference setting, but we will leave this question to other occasions.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_8,

aware of the philosophical debates about beliefs and the many options for designing appropriate logical systems.[2]

This chapter is structured as follows. in Section 8.2 we propose three different ways of defining preference in terms of priorities and beliefs. In particular, we will present a doxastic preference logic for the notion of "decisive preference" and prove an extended representation theorem for that case. Section 8.3 will extend our discussions to multi-agent case, in which we will particularly study both cooperative agents and competitive agents, and describe their characteristics in representation theorems. In Section 8.4 we move to preference over propositions, and propose a propositional doxastic preference logic. And also, we will explore the relationship between preference over objects and preference over propositions. Finally, we restate our main points in the conclusions.

8.2 Doxastic Preference Logic

8.2.1 Three Notions of Belief-Based Preference

Working with beliefs, we will first give several definitions of preference in terms of priority sequence in this section. Interestingly, the definitions we consider 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. Moreover, we consider a simpler scenario, namely, 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]

Definition 8.2 (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.

Definition 8.3 (conservative preference) Given a priority sequence of length n, two objects x and y, $Pref(x, y)$ is defined below:

---

[2] Readers who liked our plausibility models for belief in Chapters 4, 5, may also just continue thinking in these terms when reading what we have to say about the doxastic modality $B\varphi$.

[3] It would be also interesting to look at non-factual priorities containing beliefs of the agents.

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

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

Definition 8.4 (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] $:= C_1(x) \wedge \neg C_1(y)$,
$Supe_{k+1}(x, y) := Supe_k(x, y) \vee (Eq_k(x, y) \wedge C_{k+1}(x) \wedge \neg C_{k+1}(y)), k < n$,
$Supe(x, y) := Supe_n(x, y)$,
$Pref(x, y) := B(Supe(x, y))$,

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 8.1 again, but in three different variations:

A. Alice favors Definition 8.2: 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 8.3: 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 8.4: 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 [132].

Clearly, then, we have the following fact:

[4] Superiority is just defined as preference was in Chapter 7.

Fact 8.5

- *Totality holds for Definition 8.2, but not for Definition 8.3 or 8.4;*
- *Among the above three definitions, Definition 8.3 is the strongest in the sense that if Pref*(x, y) *holds according to Definition 8.3, then Pref*(x, y) *holds according to Definition 8.2 and 8.4 as well.*

It is striking that, if in Definition 8.4, 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 with $\underline{Pref}$ any more, or even $\underline{Pref}$ in terms of $Pref$ and Eq.

For all three definitions, we have the following theorem.

Theorem 8.6 $\underline{Pref}(x, y) \leftrightarrow B\underline{Pref}(x, y)$.

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. □

Corollary 8.7 $\neg\underline{Pref}(x, y) \leftrightarrow B\neg\underline{Pref}(x, y)$.

Actually, we think it is proper that Theorem 8.6 and Corollary 8.7 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.

8.2.2 Doxastic Preference Logic

If we stick to Definition 8.2, we can generalize the representation result from the previous chapter. 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.

Theorem 8.8 *The following principles axiomatize exactly the valid ones.*

(1) $\underline{Pref}(d_i, d_i)$.
(2) $\underline{Pref}(d_i, d_j) \vee \underline{Pref}(d_j, d_i)$.

(3) $\underline{Pref}(d_i, d_j) \wedge \underline{Pref}(d_j, d_k) \rightarrow \underline{Pref}(d_i, d_k)$.
(4) $\neg B \bot$.
(5) $B\varphi \rightarrow BB\varphi$.
(6) $\neg B\varphi \rightarrow B\neg B\varphi$.
(7) $\underline{Pref}(d_i, d_j) \leftrightarrow B\underline{Pref}(d_i, d_j)$.

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

Definition 8.9 (doxastic preference model) A doxastic preference model of **KD45-P** is a tuple $(S, D, R, \{\preceq_s\}_{s\in S}, V)$, where S is a non-empty set of worlds, D is a set of constants, R is a euclidean, transitive, and serial accessibility relation on S. Namely, it satisfies $\forall xyz((Rxy \wedge Rxz) \rightarrow Ryz)$, $\forall xyz((Rxy \wedge Ryz) \rightarrow Rxz)$, and $\forall x \exists y Rxy$. For each s, $\preceq_s$ 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.

Theorem 8.10 *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, transitive, 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. □

Theorem 8.11 *The logic* **KD45-P** *has the finite model property.*

Proof By standard methods. □

Theorem 8.12 (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 8.10, 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 7.9 defining the predicates $P_1, \ldots, P_n$ in a constant manner throughout the model. □

Remark 8.13 The three definitions above are not the only definitions that might be considered. For instance, we can give a variation (*) of Definition 8.3. 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 8.3, (*) is weaker in the sense that it does not require explicit positive beliefs concerning $C(x)$.

We can even combine Definition 8.2 and (*), obtaining the following:

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

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

Similarly, for Definition 8.4, if instead of $B(Supe(x, y))$, we use $\neg B\neg(Supe(x, y))$, a weaker preference definition is obtained.

8.3 Extension to the Multi-agent Case

8.3.1 Multi-agent Doxastic Preference Logic

This section extends the results of Section 8.2 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:

Definition 8.14 (reduced doxastic preference language) Let Φ be a set of propositional variables, N be a group of agents, and D be a finite domain of objects, the reduced doxastic preference language for many agents is defined in the following:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \underline{Pref}^a(d_i, d_j) \mid B^a\varphi \quad \text{where } p \in \Phi, a \in N \text{ and } d_i \in D.$$

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 $P(d_i)$.

Definition 8.15 (priority sequence for agent a) A priority sequence for agent a is a finite ordered sequence of formulas written as follows: $C_1 \gg_a C_2 \cdots \gg_a C_n (n \in \mathbb{N})$, where each C_m $(1 \leq m \leq n)$ is a formula from the language of Definition 8.14, 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.

Definition 8.16 (preference for agent a) Given a priority sequence of length n, two objects x and y, $Pref^a(x, y)$ is defined as follows:

$Pref^a_1(x, y) := B^a C_1(x) \wedge \neg B^a C_1(y)$,
$Pref^a_{k+1}(x, y) := Pref^a_k(x, y) \vee (Eq_k(x, y) \wedge B^a C_{k+1}(x) \wedge \neg B^a C_{k+1}(y)), k < n$,
$Pref^a(x, y) := Pref^a_n(x, y)$,

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

Definition 8.17 The doxastic preference logic for many agents **KD45-P$^{\mathbf{G}}$** is consists of the following principles,

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

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

Theorem 8.18 *The doxastic preference logic for many agents* **KD45-P$^{\mathbf{G}}$** *is completely axiomatized by the stated principles.*

Proof The canonical model of this logic **KD45-P$^{\mathbf{G}}$** has the required properties: The belief accessibility relation R_a is euclidean, transitive, 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. □

Theorem 8.19 *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.

Theorem 8.20 (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 $P_{n\times j+1} \gg_{a_j} P_{n\times j+2} \gg_{a_j} \cdots \gg_{a_j} P_{n\times(j+1)}$. 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. $P_{n\times j+1}$, $P_{n\times j+2}$, ..., $P_{n\times(j+1)}$ as in Theorem 7.9 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. □

8.3.2 Cooperative and Competitive Agents

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.

Theorem 8.21 (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 two agents are a and b. We now have the priority sequence $P_1 \gg_a P_2 \gg_a \cdots \gg_a P_n$, same for b. It is sufficient to show that any model $\mathfrak{M}$ 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 $\mathfrak{M}$ with a complete copy $\mathfrak{M}_E$ of $\mathfrak{M}$ 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 7.9 in E's copy E_E. What we do in the rest of $\mathfrak{M}_E$ 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 $\mathfrak{M}_E$ is bisimilar for the reduced language to the old model $\mathfrak{M}$ as is the union of $\mathfrak{M}$ and $\mathfrak{M}_E$.

Finally do the same thing for b: Add for each b-euclidean class in $\mathfrak{M}$ 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 $\neg D_m \gg_b \neg D_{m-1} \gg \cdots \gg_b \neg D_1$.

Theorem 8.22 (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 $\mathfrak{M}$ 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 $\mathfrak{M}$ with a complete copy $\mathfrak{M}_E$ of $\mathfrak{M}$ 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 $\mathfrak{M}$ with a complete copy $\mathfrak{M}_E$ of $\mathfrak{M}$ 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 $\mathfrak{M}_E$ is bisimilar for the reduced language to the old model $\mathfrak{M}$ as is the union of $\mathfrak{M}$ and all the $\mathfrak{M}_E$. □

Discussion 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, $B_a Pref_b(c, d)$ 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 cooperative agents. Moreover they show that if one wants to prove that $B_a Pref_b(c, d) \rightarrow Pref_a(c, d)$ is not valid for cooperative agents a counterexample to it in which the agents do not cooperate suffices.

8.4 Preference over Propositions

Most other authors on preference have discussed preference over propositions rather than objects. In this section, we will show that the current approach can be applied to preference over propositions as well. Following the previous section on belief-based preferences, we will propose a propositional system combining preference and beliefs. 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.

Definition 8.23 (propositional priority sequence) 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 8.2, there are various ways to do it. We are guided by the definition of decisive preference in formulating the following:

Definition 8.24 (preference over propositions) Given a propositional priority sequence of length n, we define preference over propositions ψ and θ as follows:

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

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, as in this slightly more complex definition:

Definition 8.25 A propositional priority sequence is a finite ordered sequence of sets of formulas written as follows:

$$\Gamma_1 \gg \Gamma_2 \gg \cdots \gg \Gamma_n,$$

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

A new matching definition of preference is then given by:

Definition 8.26 Given a propositional priority sequence of length n, we define preference over propositions ψ and θ as follows:

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

Remark 8.27 In fact, the priority set Γ_m could be expressed by one formula

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

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:

(1) $\underline{Pref}(\varphi, \varphi)$.
(2) $\underline{Pref}(\varphi, \psi) \wedge \underline{Pref}(\psi, \theta) \rightarrow \underline{Pref}(\varphi, \theta)$.
(3) $\underline{Pref}(\varphi, \psi) \vee \underline{Pref}(\psi, \varphi)$.
(4) $B\underline{Pref}(\varphi, \psi) \leftrightarrow \underline{Pref}(\varphi, \psi)$.
(5) $B(\varphi \leftrightarrow \psi) \rightarrow \underline{Pref}(\varphi, \psi) \wedge \underline{Pref}(\psi, \varphi)$.

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 (4) in Section 8.2. (5) 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.

Definition 8.28 (BP-model) A model of **BP** is a tuple $(S, R, \{\preceq_s\}_{s \in S}, V)$, where S is a non-empty set of worlds, R is a euclidean, transitive, and serial accessibility relation on S. Namely, it satisfies $\forall xyz((Rxy \wedge Rxz) \rightarrow Ryz)$, $\forall xyz((Rxy \wedge Ryz \rightarrow Rxz)$, and $\forall x \exists y Rxy$. Moreover, for each s, $\preceq_s$ is a quasi-linear order on propositions (subsets of S), 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.

Theorem 8.29 *The* **BP** *system is complete w.r.t the above models.*

Proof Assume $\nvdash_{\mathbf{BP}} \theta$. Take the canonical model $\mathfrak{M} = (S, R, V)$ for the formulas using only the propositional variables of θ. To each world of S 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, $\mathfrak{M}' = (S', R, V)$, which naturally is an euclidean class. Since each world in S' 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 $\mathfrak{M}'$ 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 have a representation method establishing the next result:

Theorem 8.30 (representation theorem) $\vdash_{\mathbf{BP}} \varphi$ *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

$$\Gamma_1, \ldots, \Gamma_k,$$

where Γ_1 are the best propositions ($\varphi, \psi \in \Gamma_1$ implies $\varphi \trianglelefteq \psi$ and $\psi \trianglelefteq \varphi$, Γ_i are the next best propositions, etc. Then the following is the priority sequence which results in the given order:

$$\{x \leftrightarrow \varphi \mid \varphi \in \Gamma_1\} \gg \cdots \gg \{x \leftrightarrow \varphi \mid \varphi \in \Gamma_k\}.$$

□

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, obtaining a new system $\mathbf{BP}^{\forall\exists}$. The latter has two more axioms:

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

Theorem 8.31 *The logic* $BP^{\forall\exists}$ *is complete.*

Proof By an adaption of the proof by [94]. The difference is this: [94] uses a combination of preference and the universal modality. Instead, our system is a combination of preference and belief. This means that what is preferred in our system is decided by the plausibility structure of the model. However, this does not affect Halpern's completeness proof much, and we can still use it. □

Remark 8.32 In fact, $[\leq]\varphi$ in Chapter 3 can be defined now as $\underline{Pref}(\varphi, \top)$. Then the preference used in the system $\mathbf{BP}^{\forall\exists}$ is simply the following:

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

Similarly, we get a representation-based result for this special case:

Theorem 8.33 (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.

8.4.1 Preference over Propositions and Preference over Objects

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.

Theorem 8.34 $\vdash_{\mathbf{KD45-P}} \varphi(d_1, \ldots, d_n)$ *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:

Lemma 8.35 *If* $\nvdash_{\mathbf{KD45-P}} \varphi(d_1, \ldots, d_n)$*, then for each n there is a model* $\mathfrak{M} \models \neg\varphi$ *with at least n elements.*

Proof Assume that we only have a model $\mathfrak{M} = (S, R, V)$ in which S has m elements, where $m < n$. Take one element of S, say s, and make copies of it, say, s_1, s_2,..., s_k, till we get at least n elements. If sRt, then we make s_iRt, and if tRs, then tRs_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 $\mathfrak{M}$ with only one euclidean class into a **BP** model $\mathfrak{M}'$ with at least n possible worlds in which for each s and each ψ, $\mathfrak{M}', s \models \psi(p_1, \ldots, p_n)$ iff $\mathfrak{M}, s \models \psi(d_1, \ldots, d_n)$. Let $\mathfrak{M} = (S, R, \leq, V)$, then $\mathfrak{M}' = (S', R, \trianglelefteq, V')$, where V' is like V except that for the $p_1, \ldots, p_n$, we assign $V'(p_i) = V'(p_j)$ if $d_i \leq d_j \wedge d_j \leq d_i$, otherwise,

$V'(p_i) \neq V'(p_j)$.[5] According to Lemma 8.35, there are enough subsets to do this. Finally, we set $V'(p_i) \lhd V'(p_j)$ iff $d_i < d_j$ 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, just like 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: (i) the logic over preference over all propositions as essentially richer than the logic of basic propositions or objects, or (ii) 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–though it may be of little value by itself.[6]

By applying the method of [94] we can again adapt the above proof to obtain:

Theorem 8.36 $\vdash_{\mathbf{KD45-P}} \varphi(d_1, \ldots, d_n)$ *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.

Theorem 8.37 *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

$$\varphi \rightarrow \langle \leq \rangle \psi$$

is not a local property even if $\leq$ is a subrelation of R. Nevertheless, in case $\leq$ is a subrelation of R, $B(\varphi \rightarrow \langle \leq \rangle \psi)$ does satisfy the principles of **BP** minus connectedness, and the additional $\mathbf{BP}^{\forall\exists}$ axioms, as we commented in Remark 8.32. For this purpose the following weakening of locality is sufficient:

$$\begin{aligned} \models\ & (\varphi \leftrightarrow \varphi') \wedge B(\varphi \leftrightarrow \varphi') \wedge (\psi \leftrightarrow \psi') \wedge B(\psi \leftrightarrow \psi') \\ & \rightarrow (Supe(\varphi, \psi) \leftrightarrow Supe(\varphi', \psi')). \end{aligned}$$

---

[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 $\varphi(p_1, \ldots, p_n)$.

[6] In Chapter 10, we will return to the role of structured propositions in priority graphs, showing how their "internal algebra" can be relevant to preference reasoning after all.

8.5 Conclusion

In this chapter, we have studied preference and priorities together with beliefs, as such entanglements occur naturally in real life scenarios. We constructed a new doxastic preference logic, which extended the standard logic of belief. We proved completeness and representation theorems for it, both in single-agent and multi-agent versions. This led us to consider interesting connections between preference and beliefs. Again, we strengthened the usual completeness results for logics of this kind to representation theorems. In the multi-agent case, these representation theorems were applied to cooperative and competitive agents. Finally, we proposed a new system combining beliefs and preference over propositions. To conclude this chapter, we studied the relationship between preference over objects and preference over propositions. We showed that if we think of propositional variables as representing basic propositions, then reasoning about preference over objects is the same as reasoning about preference between basic propositions.

So far, what we have explored in this part are static properties or aspects of priority-based preference, both pure and belief-entangled. In the next chapter, we will look at our earlier main concern of the dynamics of *changing preferences*, which turns out to go well with our richer modeling of the reasons underlying preference.

Chapter 9
Preference from Priorities: Dynamic Logic

9.1 Introduction

In the preceding two chapters, a rich notion of priority-based preference has been studied in stable situations. Various ways of deriving preferences from a priority sequence have been proposed, both under complete and under incomplete information. In this chapter we take up one of the main themes of this book, and address the dynamics of changes in preferences in this richer setting. What we find is that our earlier methods for plain betterness orderings generalize in an obvious manner, with a few adaptations. Therefore, this chapter will be short, since the connection, once seen, is straightforward.

As in earlier chapters, to motivate what will follow, we look at two variations of our running example of buying a house (Example 7.1):

Example 9.1 (buying with changing priorities) 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. One day, Alice luckily wins a lottery prize of ten million dollars. This changes her situation dramatically. Now she considers the quality most important, then neighborhood, then the cost.

This example shows vividly how possible changes in a priority sequence can happen, due to events that have taken place. Other sorts of changes can happen as well, for instance, something may become important that agents should take into account, or something that becomes trivial, and agents should drop it from consideration. Accordingly, in each of these cases, preference should change too. We will study all these cases in this chapter with one mechanism, which actually borrows quite a few ideas from our earlier dynamic epistemic treatment in Chapters 4 and 5.

Here is one more relevant scenario. This time, the preference change does not come from an evaluation change, but from an information change:

Example 9.2 (buying with changing beliefs) This time, Alice only considers the house's cost (C) and its 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 that $C(d_2)$ and $N(d_1)$.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_9, © Springer Science+Business Media B.V. 2011

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

Dynamic changes in Alice's preference are now due to her belief change. For belief-based preference, this is easy to understand: When beliefs change, so do our preferences.

In other words, both changes in priority sequence and changes in belief can cause preference change. In this chapter we study both.[1]

This chapter is organized as follows. In Section 9.2 we will focus on priority changes, and the preference changes which they cause. Section 9.3 will deal with belief change that leads to preference change, and we will look at both cases of Chapters 4, 5: Belief changes due to "hard information", and belief changes due to "soft information". Taking the *DEL* approach to our priority structures, we treat the relevant events as dynamic actions and find their matching reduction axioms. Our conclusions from all this are in Section 9.4.

9.2 Preference Change due to Priority Change

In this section, 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 7.2. Then we write $Pref_{\mathfrak{C}}(x, y)$ for the preference defined from that priority sequence.

Let us consider the following possible operations:

(1) $\mathfrak{C}^\frown C$ for adding C to the right of $\mathfrak{C}$,
(2) $C^\frown \mathfrak{C}$ for adding C to the left of $\mathfrak{C}$,
(3) $\mathfrak{C}^-$ for the sequence $\mathfrak{C}$ with its final element deleted,
(4) $\mathfrak{C}^{i \leftrightarrows i+1}$ for $\mathfrak{C}$ with its ith and i+1th priorities switched.

Those operations reflect possible changes to the priority sequence. The first one and the second one introduce something new into the sequence, as the least important element and the most important element, respectively. The third one deletes the existing least important element. The final one shifts the importance of the adjoining two elements of the priority sequence. In terms of the preference before and after such changes, it is clear that we have the following relationships:

$$Pref_{\mathfrak{C}^\frown C}(x, y) \leftrightarrow Pref_{\mathfrak{C}}(x, y) \vee (Eq_{\mathfrak{C}}(x, y) \wedge C(x) \wedge \neg C(y)),$$
$$Pref_{C^\frown \mathfrak{C}}(x, y) \leftrightarrow (C(x) \wedge \neg C(y)) \vee ((C(x) \leftrightarrow C(y)) \wedge Pref_{\mathfrak{C}}(x, y)),$$
$$Pref_{\mathfrak{C}^-}(x, y) \leftrightarrow Pref_{\mathfrak{C},n-1}(x, y),$$

[1] Note that priority change leads to a preference change in a way similar to "entrenchment change" in belief revision theory (see [162]). Still, we stick to the methodology of dynamic epistemic logic, now applied to belief.

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

These relationships enable us to describe preference change due to changes of the priority sequence in the manner of dynamic epistemic logic. To do that, we introduce the following four dynamic actions: $[^{+}C]$ 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 ith and i+1th 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 "suggestion C" $[\sharp C]$ in Chapter 3. We will discuss more connections of this sort in Chapter 10.

In the context of incomplete information when we have the language of belief, we can obtain similar reduction axioms for Definitions 8.2 and 8.3. For instance, for Definition 8.2, we need only replace C by BC and $\neg C$ by $\neg BC$. For Definition 8.4, 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.

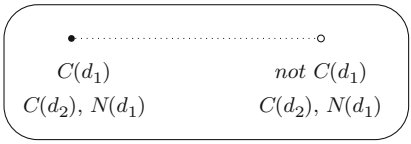
9.3 Preference Change due to Belief Change

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 [15] and [29] 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.

9.3.1 Hard Information

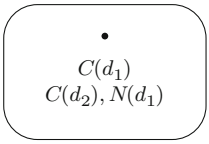
Consider the Example 9.2 again, let us assume that Alice treats the information obtained as hard information. She simply adds new information to her stock of beliefs. Figure 9.1 shows the situation before Alice's reading the newspaper.

Fig. 9.1 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 Fig. 9.2. The $\neg C(d_1)$-world is eliminated from the model: Alice has updated her beliefs. Now she prefers d_1 over d_2.

Fig. 9.2 Updated model



We have assumed that we are using the elimination semantics (e.g. [27, 73], 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 [29], 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 Chapter 8 by B^{ψ}. Assuming A is a formula without belief operators, an easy calculation gives us another form of the reduction axiom:

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

This concludes our discussion on preference change caused by belief change when information is fully accepted.

9.3.2 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 $\preceq$. 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, $w \prec u$. Otherwise $v \prec v'$ 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 [192] and [163], adopted by [29] 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 [29] 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 then the following:

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

By the fact that we have these reduction axioms here, the completeness analysis in [29] for dynamic logics of belief can be extended straightforwardly to a dynamic preference logic. We will not spell out the details here.

9.4 Conclusion

This chapter explored how to model dynamic changes of preference within the framework of reason-based preference. In line with the various ways of defining preference proposed in the previous two chapters, two main cases have been considered. The preference change may be due to a change in the priority base, or it may be

caused by a belief change. For the first case, we have looked at four possible natural operations on the priority base that transform agents' priorities. For the later case, we studied belief change under hard information, and under soft information, applying the recent *DEL* approach to this area – and we found complete reduction axioms for those cases. Our central example was the operation of "radical" lexicographic belief upgrade, and we showed how earlier results on belief revision in Chapters 4, 5 can be adopted smoothly to the richer setting with priority-based preference. As we shall see later in Chapter 10, these results are not confined to linear priority orders and connected betterness relations: They generalize easily to partial priority orders and betterness pre-orders.

The fundamental new contribution of this part of our book is that, for the first time, reasons for preference have been included in a systematic logical account. In doing so, we provided a formalization for the notion of extrinsic preference proposed by von Wright. All this suggests that, in order to truly understand preference, a more structured setting is called for. In the next Part V of this book, we will combine the results achieved in Part III with those in Part IV, and elaborate this richer view of preference in greater logical detail. Following that, in Part VI of this book, we will show how this combination also makes sense in major areas of application, such as deontics and game theory.

Part V
A Two-Level Perspective on Preference

Chapter 10
A Two-Level Perspective on Preference

10.1 Introduction

In Part III of this book,we presented a modal logic approach to preference and preference change via betterness relations. Then, in Part IV, we developed what might be seen as a competing priority-based view of preference. These two perspectives had different intuitions, both plausible and attractive. Even so, the question naturally arises how the two are related. The aim of the present chapter is to draw a comparison, connect them, and try to integrate them.

Clearly, despite the difference in starting point, the agendas and methods of Part III and Part IV are very similar. There is preference structure which eventually shows in possible worlds models, there are matching logical languages, and there are dynamic actions that change preferences in a systematic manner. Can we have the benefits of both? The purpose of this Chapter is to develop a *two-level perspective* on preference models and preference change, that one can argue for from a conceptual, but also a technical point of view. Conceptually, our main point is that von Wright's distinction between "intrinsic" and "extrinsic" preference, discussed in Chapter 1, leaves both kinds natural as an aspect of human agency. We do not want to choose between them, we want to be able to have both, and perhaps even switch between them as the occasion arises. Technically, our main consideration is simply this: A combination of the two views is quite feasible, and indeed, it makes us see a number of interesting new logical questions. That, too, is a good reason for embarking on a merge of our two preference levels.

To achieve all this, we will combine the systems of Chapters 3 and 7, putting betterness and priority structure together. While this works well, we also get some immediate issues. Notably, in this "two-level view", there will then be two different logics for preference change: one as dynamics of changing betterness orders, and another as priority dynamics of changing reasons. In this chapter, we will study their correlations in some technical detail. Our analysis finds some promising correspondence results between the two levels, but we will also show that there is no simple algorithm reducing changes at one level to those at the other. Thus, we conclude once more that both levels are needed, also from a dynamic perspective. But we do end with some thoughts on how the two levels can transform into each other, using a technical representation result that will be explained below.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_10,

More precisely, this chapter is organized as follows. Section 10.2 then merges ideas from [117] and [6], extending the linear priority sequence to partial ordered priority graphs. In particular, we extend the main representation theorem in [117] for priority-induced total betterness relations to pre-orders, representing the more realistic case of possibly incomparable or even conflicting priorities. Now changes can also take place at the level of priorities, and Section 10.3 studies its basic dynamics.

Next we discuss what logics are naturally supported by these priority structures, and Section 10.4 explores, in particular, an interesting new interplay between an "external" language of graph composition and the "internal" language of prioritized descriptive propositions inside graphs.

We then turn to connections between the two paradigms. Given the occurrence of dynamic changes at the two levels, Section 10.5 investigates their connections, finding a general inclusion result from basic priority changes to *PDL*-definable betterness transformers. But also, we find examples that show that there is no inclusion, either way. Section 10.6 then discusses what this means for the interplay of extrinsic and intrinsic preferences, in terms of logics and languages, though we do not propose any large merged system. Instead, we point out some new issues about preference that arise here, different from the existing literature, such as the fundamental role of dynamic operations of *language change* between the two levels.

We conclude that having both levels around in preference logic is important for two reasons. First, it allows for more realistic and sophisticated modeling of preference scenarios – a theme developed at greater length for natural language and general preference in [131], and for deontic reasoning in Chapter 11 below. But at the same time, we show that the two-level view also has intrinsic attractions, being a source of interesting new notions and logical problems.

10.2 An Extension to Priority Graphs

The idea of introducing priority in preference logic as discussed here comes from [117] (explained in Chapter 7), in which betterness order of objects or worlds is derived from a linearly ordered priority graph of "important propositions", viewed as properties of worlds or objects. However, we take a more general approach here, since an assumption of linear order seems unrealistic for the priorities that determine most of our preferences. Our main structures will be strict partial orders of propositions ("directed acyclic graphs"). In what follows, we start with some basic definitions, borrowing ideas from the seminal paper [6] that allow for partially ordered priorities and pre-ordered betterness relations.

10.2.1 Priority Graphs and Extrinsic Betterness

The following definitions contain a "free parameter" for a *language L* that can be interpreted in the earlier modal betterness models. For simplicity, we will take this

to be a simple propositional language of properties – though generalizations are possible.

Definition 10.1 (priority graph) A priority graph $\mathcal{G} = (P, <)$ is a strictly partially ordered set of propositions in a language L.

Here is how one derives a betterness relation from a priority graph.

Definition 10.2 (betterness from a priority graph) Let $\mathcal{G} = (P, <)$ be a priority graph, and $\mathfrak{M}$ a model in which the language L defines properties of objects. The induced betterness relation $\leq_{\mathcal{G}}$ is defined as follows:

$$y \leq_{\mathcal{G}} x := \forall P{\in}\mathcal{G}((Py \rightarrow Px) \vee \exists P'{<}P(P'x \wedge \neg P'y)).$$

This is best understood as follows. In principle, $y \leq_{\mathcal{G}} x$ requires that x has every property in the graph that y has. But there is a possibility of "compensation": In case y has P while x does not, this is still admissible, provided that there is some property P' with higher priority in the graph where x does better: x has P' while y lacks it.[1]

Remark 10.3 (on notation for propositions) For reasons of readability, we switch notation here, and use capital letters P, A and B for propositions that occur in priority graphs, rather than Greek letters φ, ψ.

For *totally ordered graphs* $\mathcal{G}$, graph-induced betterness order reduces to the usual lexicographic ordering (for details, cf. [131]):

$$y \leq^{lin}_{\mathcal{G}} x := \forall P{\in}\mathcal{G}(Px{\leftrightarrow}Py) \vee \exists P'{\in}\mathcal{G}(\forall P{<}P'(Px{\leftrightarrow}Py) \wedge (P'x{\wedge}{\neg}P'y)).$$

At an opposite extreme, *flat priority graphs* have no links between different propositions. In that case, our definition reduces to an order known from the area of default reasoning (cf. [192]):

$$y \leq^{flat}_{\mathcal{G}} x := \forall P{\in}\mathcal{G}(Py \rightarrow Px).$$

We refer to [6] for many mathematical properties of priority graphs, a few of which will be used later. For now, just note that, given a priority graph $\mathcal{G}$ and a derived model $\mathfrak{M}_{\mathcal{G}} = (S, \leq_{\mathcal{G}}, V)$, the latter may be viewed as a reason-based betterness model in the sense discussed in Chapter 3.

10.2.2 An Extended Representation Theorem

The relations induced by priority graphs in betterness models have various special features. For instance, it is easy to see the following:

[1] There are a few differences here with the general approach in [6] that we do not spell out. In particular, to make things comparable, note that each unary property P is naturally associated with the binary relation P such that for all x and y, xPy iff Px implies Py.

Fact 10.4 *Priority graphs induce pre-orders in which betterness relations hold uniformly between whole "zones" of the model consisting of all worlds satisfying the same propositions in our graph language.*

This observation points the way to a more general issue. We have now described how preference orders arise extrinsically from a priority structure. A natural converse question is the following. If we start with a given betterness order, can we always find some priority graph from which the given order is derived? The answer is the following representation result, generalizing an earlier one for total orders from [117]:

Theorem 10.5 *Let $\mathfrak{M} = (S, \leq, V)$ be any modal model, without constraints on its relation. The following two statements are equivalent:*

(a) The relation $\leq$ is a reflexive and transitive order.
(b) There is a priority graph $\mathcal{G} = (P, <)$ such that, for all worlds x, $y \in S$, $y \leq x$ iff $y \leq_{\mathcal{G}} x$.

Proof In the direction from (b) to (a), it is easy to see from Definition 10.2 that all relations $\leq_{\mathcal{G}}$ induced by priority graphs are pre-orders.

Conversely, consider the direction from (a) to (b).We first define a "cluster" as a maximal subset X of S such that $\forall y, z \in X$: $y \leq z$. Clusters exist by Zorn's Lemma, and different clusters are disjoint by their maximality. Each point x of the model $\mathfrak{M}$ belongs to a cluster, which we call C_x. Next, we define a natural ordering of clusters reflecting that of the worlds.

$C' \trianglelefteq C$ if $\exists y \in C', \exists x \in C$: $y \leq x$.

Next, we prove a connection with the given pre-order among objects:

Lemma 10.6 $y \leq x$ *iff* $C_y \trianglelefteq C_x$.

Proof ($\Rightarrow$) By definition, $x \in C_x$ and $y \in C_y$, so $C_y \trianglelefteq C_x$.
($\Leftarrow$) If $C_y \trianglelefteq C_x$, then by definition $\exists u \in C_x, v \in C_y$ with $v \leq u$. So $u \leq x$ (since $x \in C_x$) and $y \leq v$ (since $y \in C_y$) – and by transitivity, we also have $y \leq x$. □

Now, consider the set of all clusters, viewed as nodes in a graph. We impose an order of greater priority in the upward direction of the preceding cluster order.[2] Note that this is not just a pre-order, but a strict acyclic partial order, since we have identified objects in equivalence classes. Therefore, we indeed have a priority graph $\mathcal{G}^\bullet$ in the sense of the above definition.

Now we need to show that the relation induced by the graph $\mathcal{G}^\bullet$ matches up with the given betterness relation $\leq$ in the model $\mathfrak{M}$:

Lemma 10.7 $y \leq x$ *iff* $y \leq_{\mathcal{G}^\bullet} x$.

[2] This choice of direction is just a convention – but we need to fix one in our proof.

From left to right, assume that $y \leq x$. We must show that $y \leq_{\mathcal{G}^\bullet} x$, i.e., $\forall P \in \mathcal{G}^\bullet((Py \rightarrow Px) \vee \exists P' < P(P'x \wedge \neg P'y))$. So, consider any cluster proposition P. If $Py \rightarrow Px$, we are done. So, let $Py \wedge \neg Px$. Then we have $P = C_y$, since as before, our cluster propositions form a disjoint partition. Moreover, we have $C_y \trianglelefteq C_x$, by Lemma 10.6 applied to $y \leq x$. But then, since $C_y \neq C_x$ (they are disjoint, as x is not in C_y), we get $C_y \triangleleft C_x$. It is easy to see that C_x is the "compensating" property P' for x that we need for an appeal to the second disjunct in Definition 10.2.

From right to left, let $y \leq_{\mathcal{G}^\bullet} x$. Consider the predicate $P = C_y$, for which Py holds. First assume that Px. Since P $(= C_y)$ is a cluster, we have $y \leq x$. Next, let $\neg Px$. By the "compensation clause" of $y \leq x$, it follows that $\exists P' < P$: $P'x \wedge \neg P'y$. Clearly, this predicate P' can only be C_x, and so $C_y \triangleleft C_x$, $C_y \trianglelefteq C_x$, and by Lemma 10.6 once more, we get $y \leq x$. □

This representation theorem may be viewed in several ways. It tells us that the general logic of derived extrinsic betterness orderings is still just that of the earlier pre-orders. But it also tells us that any pre-order can be "rationalized" as an extrinsic reason-based one without disturbing the model as it is. We will return to this second perspective in Section 10.4 below.

10.3 Basic Operations on Priority Graphs

A priority graph as a structured set of propositions naturally suggests dynamic changes. New properties may become important to agents' preferences, old ones may become less important, or totally irrelevant. In this section, we introduce some basic update operations on priority graphs.

10.3.1 Basic Graph Update

We start with very simple changes, represented in an algebraic format. These involve the following two basic operations from [6]. Given any two priority graphs $\mathcal{G}, \mathcal{G}'$,

- the *sequential composition* $\mathcal{G};\mathcal{G}'$ adds the graph $\mathcal{G}$ on top of $\mathcal{G}'$ in the order: All nodes in the first come before all those in the second,
- the *parallel composition* $\mathcal{G} \| \mathcal{G}'$ is the disjoint union of the graphs $\mathcal{G}$ and $\mathcal{G}'$, without any order links between them.

One can either think about syntactic operations on actual graphs here, or about corresponding terms in an algebraic formalism: We refer to the above-cited paper for details of the match. What follows can be understood while freely switching between these two perspectives.

Give a priority graph $\mathcal{G}$, there are obvious options for placing a new proposition A. One can make it the highest priority, or the lowest, or one may also rank it just

side by side with $\mathcal{G}$. Generalizing this placement of old and new material in a current graph suggest the following notion:

Definition 10.8 (basic graph updates) Let "A" stand for the priority graph with one single node A. The set $\alpha(\mathcal{G}, A)$ of *basic graph updates* is defined by the following inductive syntax rule:

$$\alpha(\mathcal{G}, A) := A \mid \mathcal{G}_1; \mathcal{G}_2 \mid \mathcal{G}_1 \| \mathcal{G}_2.$$

Basic graph updates are about the simplest syntactic ways of combining a given priority graph $\mathcal{G}$ with a new proposition A.

But clearly, there are other operations that change priority graphs. An obvious counterpart to insertions are *deletions*, acts of "dropping an issue":

Definition 10.9 (top deletion) The operation of *top deletion* on a (non-empty) priority graph $\mathcal{G}$ deletes all propositions that are not dominated by another in the graph order, leaving the rest in its old order. We write $del(\mathcal{G})$ for the result of a top deletion.

Still further operations would permute propositions in graphs, changing relative importance (Chapter 9, [83]), or insert new propositions in intermediate positions. But the present examples will suffice for showing the dynamics of what may be called "priority management".

10.3.2 Graph Algebra

Describing the dynamics at the present level invites algebraic analysis in a language of terms and identities.

Here are some basic algebraic laws that hold for *sequential* and *parallel composition* of graphs from [6][3]:

Fact 10.10 *The following laws hold for graph-induced betterness relations:*

(1) $\leq_{\mathcal{G}_1 \| \mathcal{G}_2} = \leq_{\mathcal{G}_1} \cap \leq_{\mathcal{G}_2}$.
(2) $\leq_{\mathcal{G}_1; \mathcal{G}_2} = (\leq_{\mathcal{G}_1} \cap \leq_{\mathcal{G}_2}) \cup <_{\mathcal{G}_1}$.

Based on these facts, we sometimes write a short-hand $\mathcal{G}_1 \| \mathcal{G}_2 \equiv \mathcal{G}_1 \cap \mathcal{G}_2$ and also $\mathcal{G}_1; \mathcal{G}_2 \equiv (\mathcal{G}_1 \cap \mathcal{G}_2) \cup \mathcal{G}_1^{<}$ when matters are clear in context.

Next, call two priority graphs *equivalent* (denoted as $\equiv$) if they induce the same relation in every model. The same notion makes sense for algebraic terms, viewed as describing priority graphs, and validity then has the obvious meaning.

As an illustration, we list a few simple algebraic properties that are validated by the above definitions, as well as two useful identities for strict relations that will return in Section 10.5 below:

---

[3] Again, these should be read keeping in mind the close connection between priority graphs and algebraic terms formed with the operations ; , $\|$.

Fact 10.11 *The following algebraic equations are valid:*

(1) $\mathcal{G};\mathcal{G} \equiv \mathcal{G}$.
(2) $\mathcal{G} \parallel \mathcal{G} \equiv \mathcal{G}$.
(3) $\mathcal{G}_1 \parallel \mathcal{G}_2 \equiv \mathcal{G}_2 \parallel \mathcal{G}_1$.
(4) $(\mathcal{G}_1 \parallel \mathcal{G}_2)^{<} \equiv (\mathcal{G}_1^{<} \parallel \mathcal{G}_2) \cup (\mathcal{G}_1 \parallel \mathcal{G}_2^{<})$.
(5) $(\mathcal{G}_1;\mathcal{G}_2)^{<} \equiv (\mathcal{G}_1^{<} \cup (\mathcal{G}_1 \parallel \mathcal{G}_2^{<}))$.

We leave the simple proofs of these identities to the reader.

10.4 Logics for Priority and Extrinsic Preference

Like the earlier betterness structure of possible worlds, priority structure invites simple logical languages that can bring out key features of reasoning. We will define a few of these, linking up with the existing literature, and suggesting a number of new topics that deserve attention. But our aim in this chapter are not the logical systems per se. We discuss them mainly as additional evidence for the naturalness of making priority structure explicit.

10.4.1 Modal Logic of Graph-Induced Betterness

The simplest way of putting priority structure into a modal language of the sort we have used many times before for betterness models is found in [83]. We introduce modalities that are labeled by priority graphs inducing the accessibility relations. Moreover, in this setting, it is useful to add a device from hybrid modal logic, viz. *nominals*: special proposition letters that are true at one world only (cf. [9] for a full account of this device).

Definition 10.12 (modal graph language) Let Φ be a set of propositional variables with $p \in \Phi$, and $\mathtt{Nom}$ a set of nominals with $n \in \mathtt{Nom}$. Let $\mathbb{G}$ be a set of graphs with the variable $\mathcal{G}$ ranging over $\mathbb{G}$. The modal graph language is then defined by the following syntax rule:

$\varphi := n \mid p \mid \neg\varphi \mid \psi \wedge \varphi \mid \langle\mathcal{G}\rangle\varphi \mid \langle\mathcal{G}\rangle^{<}\varphi \mid E\varphi$.
$\mathcal{G} := \mathcal{G}_1;\mathcal{G}_2 \mid \mathcal{G}_1 \parallel \mathcal{G}_2$.

If we drop the graph symbols, we get the basic betterness modalities. The semantic structures described by the modal graph language may be viewed as having families of extrinsic betterness relations, with the priority graphs supplying the reasons for the preference.[4]

Definition 10.13 (modal graph model) A modal graph model is a tuple $\mathfrak{M} = (S, \mathbb{G}, \leq_{\mathcal{G}}, V)$, where S is a non-empty set of possible worlds, $\mathbb{G}$ a set of graphs, $\leq_{\mathcal{G}}$

[4] We will discuss possible co-existence of extrinsic graph-based with intrinsic primitive betterness relations in Section 10.4.2 below.

a family of betterness relations induced by graphs $\mathcal{G} \in \mathbb{G}$, and $V: \Phi \cup \mathtt{Nom} \to \mathcal{P}(S)$ is a valuation assigning sets of worlds to propositional variables, and singleton sets of worlds to members of Nom.

Definition 10.14 (truth definition) Given a modal graph model $\mathfrak{M}$, the truth definition for formulas is as follows, omitting obvious Boolean cases:

(1) $\mathfrak{M}, s \models n$ iff $\{s\} = V(n)$.
(2) $\mathfrak{M}, s \models p$ iff $\{s\} \in V(p)$.
(3) $\mathfrak{M}, s \models \langle \mathcal{G} \rangle \varphi$ iff for some t with $s \leq_{\mathcal{G}} t$, $\mathfrak{M}, t \models \varphi$.
(4) $\mathfrak{M}, s \models \langle \mathcal{G} \rangle^{<} \varphi$ iff for some t with $s <_{\mathcal{G}} t$, $\mathfrak{M}, t \models \varphi$.

Now, given the analysis of graph-induced relations in Section 10.4, complex relations $\leq_{\mathcal{G}}$ can be recursively reduced to basic betterness relations using the graphs identities stated in Fact 10.10:

$$\mathcal{G}_1 \| \mathcal{G}_2 \equiv \mathcal{G}_1 \cap \mathcal{G}_2 \quad \text{and} \quad \mathcal{G}_1; \mathcal{G}_2 \equiv (\mathcal{G}_1 \cap \mathcal{G}_2) \cup \mathcal{G}_1^{<}.$$

As for the logic of this language of extrinsic preference – at least when the graphs are finite, [83] presents a complete axiomatization. Its main axioms follow the identities formulated in Facts 5.3 and 5.4:

$$\langle \mathcal{G}_1 \| \mathcal{G}_2 \rangle n \leftrightarrow \langle \mathcal{G}_1 \rangle n \wedge \langle \mathcal{G}_2 \rangle n.$$
$$\langle \mathcal{G}_1 \| \mathcal{G}_2 \rangle^{<} n \leftrightarrow (\langle \mathcal{G}_1 \rangle^{<} n \wedge \langle \mathcal{G}_2 \rangle n) \vee (\langle \mathcal{G}_1 \rangle n \wedge \langle \mathcal{G}_2 \rangle^{<} n).$$
$$\langle \mathcal{G}_1; \mathcal{G}_2 \rangle n \leftrightarrow (\langle \mathcal{G}_1 \rangle n \wedge \langle \mathcal{G}_2 \rangle n) \vee \langle \mathcal{G}_1 \rangle^{<} n.$$
$$\langle \mathcal{G}_1; \mathcal{G}_2 \rangle^{<} n \leftrightarrow (\langle \mathcal{G}_1 \rangle n \wedge \langle \mathcal{G}_2 \rangle^{<} n) \vee \langle \mathcal{G}_1 \rangle^{<} n.$$

These axioms allow one to reduce complex priority relations to simple ones, after which the whole language reduces to the modal logic of weak and strict atomic betterness orders (cf. Chapter 4). In particular, this modal graph logic encodes the graph algebra of [6], while it also remains decidable.[5]

10.4.2 Internal Versus External Graph Language

The preceding language describes graph-induced betterness structure at worlds. However, it does not describe the equally natural "internal language" L of the graphs, that we used to define priority graphs in the first place. Confusing the two may lead to curious phenomena:

Example 10.15 (dangers of self-reference) If assertions inside graphs can contain modalities for the induced strict betterness order $<$, then we can have a graph consisting of just the proposition $A = \langle < \rangle \top$ (alone, and hence in top position) saying that "there is a strictly better world". But this gives a sort of Liar Paradox: A world satisfying this A should be maximal in the induced order, but by the definition of A it cannot be maximal!

[5] There is a translation into the decidable *two-variable fragment* of first-order logic.

Finding "harmless" modal extensions for the purely propositional internal graph language seems an interesting open problem. In what follows we stick with the basic case, emphasizing other issues.

An evident next task is an extension of our system for valid reasoning. We already have the general graph algebra for ; and ‖ of [6]: see Section 10.3.2 above. But the internal language also generates further valid principles. For instance, one can replace propositions in priority graphs by logical equivalents without change in their induced relations. Here is a general fact:

Theorem 10.16 *The modal logic encoding both the external and the internal graph algebra is completely axiomatizable.*

Proof This can be shown as follows. We already have the above complete modal logic for the external algebra of ; and ‖. The only thing we need to add is a description of specific propositions in graphs. And given the earlier reduction argument for complex graphs, it suffices to just give a valid axiom for graphs consisting of a single proposition A. Here it is:

$$\langle A\rangle\varphi \leftrightarrow (E(A \wedge \varphi) \vee (\neg A \wedge E\varphi)).$$

It is easy to check that this equivalence valid. □

The attraction of this logic shows in concrete graph transformations:

Example 10.17 (distributive normal form under priority) The graph $A; B$ is equivalent to the graph $A \wedge B; A \wedge \neg B; \neg A \wedge B; \neg A \wedge \neg B$. This can be shown by computing the relations in both cases using the above special axiom (Fig. 10.1).

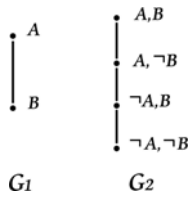


Fig. 10.1 Two equivalent graphs

In particular, the proof of Theorem 4.4 in our earlier Chapter 4 suggests a general principle that holds in this internal graph algebra. Its representation method for preorders introduced a graph of propositions that formed a disjoint *partition* of the whole domain of worlds. Likewise, using the above completeness theorem, one can prove:

Fact 10.18 *Each priority graph has an equivalent graph whose propositions form a partition of the logical space.*[6]

[6] The main idea is that, like with distributive normal forms in propositional logic, one can effectively transform any given priority graph into an equivalent one where the propositions are complete conjunctions of literals for all relevant proposition letters.

Many other language issues make sense for priority structures, but the themes introduced here may suffice to show their interest.[7] By making graphs explicit, we have replaced dynamics by statics, since talking about a graph update will now match earlier reduction axioms. For instance, $\langle A; \mathcal{G}\rangle\varphi \leftrightarrow \langle\Uparrow\rangle\langle\mathcal{G}\rangle A$.

Also, finding a logic for *deleting items* from a graph is an open problem in [83]. But it depends on how one casts the language. E.g., restricting attention to linear graphs, add an operator 'head' from graphs to their topmost proposition, and an operator 'tail' producing the rest of the graph. Now we do get a logic of deletion, with obvious valid principles such as $\langle\mathcal{G}\rangle\varphi \leftrightarrow \langle head(\mathcal{G}); tail(\mathcal{G})\rangle\varphi$.

More syntactic views of priority graphs Our final point is very different in thrust. The above logics all view priority graphs from a semantic perspective, by focusing on their derived betterness orders. But it is also very natural to view priority graphs as syntactic objects where syntactic manipulations can have effects of their own, even when they do not change the induced betterness order. This more fine-grained view lies behind the syntactic "agenda dynamics" of [39]. A syntactic change in a graph may be viewed as an intensional change in the presentation of one's reasons for preference, akin to performing an inference step.[8] Indeed, fine-grained deductive exploration of the effects of priority structure occurs, for instance, in much legal reasoning (cf. [90]). We think that adding such a more syntactic view of preference structure, with a matching more fine-grained graph dynamics makes sense, but leave this for another occasion.

While language design is not the main topic of this chapter, this Section will have illustrated that the priority level suggests much richer "preference logics" than the usual formalisms going by that name.

10.5 Relating Betterness and Priority Dynamics

Having completed our development of priority-based preference logic, the question arises how graph update matches with the betterness dynamics in Chapter 4. To study the connection, we first introduce a technical notion:

Definition 10.19 Let α: $(\mathcal{G}, A) \rightarrow \mathcal{G}'$, with $\mathcal{G}$, $\mathcal{G}'$ priority graphs, and A a new proposition. Let σ be a map from $(\leq, A)$ to $\leq'$, where $\leq$ and $\leq'$ are betterness relations over worlds. We say that α induces σ, if always:

$$\sigma(\leq_{\mathcal{G}}, A) \quad = \quad \leq_{\alpha(\mathcal{G},A)}$$

We call *the operation* α *PDL-definable* if it induces a relation transformer σ that is *PDL*-definable in the format of Chapter 4.

---

[7] One obvious connection is with the *dynamic betterness logics* of Chapter 4.

[8] Reference [190] extends current semantic dynamic epistemic logic to systems that can deal with syntactic acts of inference.

10.5.1 Cases of Harmony

We first look at two uniformly definable cases where priority dynamics is a perfect match with betterness dynamics. We start with our pilot example for betterness change from Chapter 4, that of "suggestions".

Fact 10.20 *Taking a suggestion A given some betterness relation over worlds is induced by the following basic graph update at the priority level: $\mathcal{G}\|A$. More precisely, the following diagram commutes:*

$$\begin{array}{ccc} (\mathcal{G}, <) & \xrightarrow{\|A} & ((\mathcal{G}\|A), <) \\ \downarrow & & \downarrow \\ (W, \leq) & \xrightarrow{\sharp A} & (W, \sharp A(\leq)) \end{array}$$

Proof We need to prove the following equivalence:

$$y \leq_{\mathcal{G}\|A} x \quad \text{iff} \quad y \,\sharp A(\leq_{\mathcal{G}})\, x.$$

($\Leftarrow$) After $\sharp A$ the relation between y and x becomes this:

$$\sharp A(\leq_{\mathcal{G}}) := (?A; \leq_{\mathcal{G}}; ?A) \cup (?\neg A; \leq_{\mathcal{G}}; ?\neg A) \cup (?\neg A; \leq_{\mathcal{G}}; ?A).$$

In terms of a relation between arbitrary worlds x and y, the above three cases give the implication $Ay \to Ax$. By $y \leq_{\mathcal{G}} x$, we also have that $\forall P \in \mathcal{G}$: $Py \to Px$. Hence we get $\forall P \in \mathcal{G}\|A((Py \to Px) \vee \exists P' < P\ (P'x \wedge \neg P'y))$, and this is precisely what $y \leq_{\mathcal{G}\|A} x$ says.

($\Rightarrow$) Let $y \leq_{\mathcal{G}\|A} x$, that is, $\forall P \in \mathcal{G}\|A((Py \to Px) \vee \exists P' < P\ (P'x \wedge \neg P'y))$. In particular, it cannot be the case that $Ay \wedge \neg Ax$. Thus, out of all pairs in the given relation R, those satisfying $(?A; \leq_{\mathcal{G}}; ?\neg A)$ can no longer occur. This is precisely how we defined the relation $y \,\sharp A(\leq_{\mathcal{G}})\, x$. □

Simple as it is, this argument shows how natural order-changing operations at both levels of our preference models can be tightly correlated.

Next, consider a priority graph $(\mathcal{G}, <)$, with a new proposition A added on top. The dynamics at the two levels is now correlated as follows:

Fact 10.21 *Prefixing a new proposition A to a priority graph $(\mathcal{G}, <)$ induces the radical upgrade operation $\Uparrow A$ on possible worlds models. More precisely, the following diagram commutes:*

$$\begin{array}{ccc} (\mathcal{G}, <) & \xrightarrow{A;\mathcal{G}} & ((A; \mathcal{G}), <) \\ \downarrow & & \downarrow \\ (W, \leq) & \xrightarrow{\Uparrow A} & (W, \Uparrow A(\leq)) \end{array}$$

Proof Again, we have to prove a simple equivalence:

$$y \leq_{A;\mathcal{G}} x \quad \text{iff} \quad y \Uparrow A(\leq_{\mathcal{G}}) x.$$

($\Leftarrow$) After the operation $\Uparrow A$, the relation between y and x is:

$$\Uparrow A(\leq_{\mathcal{G}}) := (?A; \leq_{\mathcal{G}}; ?A) \cup (?\neg A; \leq_{\mathcal{G}}; ?\neg A) \cup (?\neg A; \top; ?A)$$

Call these disjuncts (a), (b) and (c), respectively. We prove $y \leq_{A;\mathcal{G}} x$, i.e.:

$$\forall P \in A; \mathcal{G} : ((Py \to Px) \vee \exists P' < P(P'x \wedge \neg P'y))$$

With disjunct (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 $\mathcal{G}$. With disjunct (c), since $(Ax \wedge \neg Ay)$, for any pair of y and x, A is the compensating predicate G' in $A; \mathcal{G}$ that we need for Definition 10.2. Thus in all cases, we have $y \leq_{A;\mathcal{G}} x$.

($\Rightarrow$) Let $y \leq_{A;\mathcal{G}} x$. There are two cases. (i) For all $P \in A; \mathcal{G}$ $(Py \to Px)$. In particular then, $Ay \to Ax$, and we get disjuncts (a) and (b). (ii) For all $P \in A; \mathcal{G}$, there is a $P' \in A; \mathcal{G}$ such that $(P'x \wedge \neg P'y)$. Then $P' = A$ or $P' \in \mathcal{G}$. If $P' = A$, $Ax \wedge \neg Ay$, and we get disjunct (c). If $P' \in \mathcal{G}$, then, by the prefixing, $A < P'$, by assumption we have $Ay \to Ax$, and again we get truth of the disjuncts (a) and (b).

□

These perspicuous diagrams show that the two kinds of dynamics can dovetail very well, reinforcing our general idea of two level preference.

10.5.2 General Connections

We now develop this theme in more generality. For a start, one can think of the preceding connections in two different ways: from priority-level transformers to matching world-level relation transformers, or vice versa, from world-level transformers to inducing priority-level transformers. We will show that some positive matches hold, explaining our examples.

Theorem 10.22 *Basic graph updates induce PDL-betterness transformers.*

Proof There are various ways to prove this result, that have independent interest, but here we only give one by brute enumeration:

Lemma 10.23 *All basic graph updates reduce to a finite set of cases.*

Up to graph equivalence, all basic graph updates reduce to the 5 cases listed in Tables 10.1 and 10.2. These two tables show closure under the two basic operations; and $\|$, respectively:

We just do a few cases as an example, using our earlier observations in Section 10.3 on valid principles for the algebra of equivalent priority graphs:

Table 10.1 Sequential composition

;	A	$\mathcal{G}$	$A; \mathcal{G}$	$\mathcal{G}; A$	$A \| \mathcal{G}$
A	A	$A; \mathcal{G}$	$A; \mathcal{G}$	$A; \mathcal{G}$	$A; \mathcal{G}$
$\mathcal{G}$	$\mathcal{G}; A$	$\mathcal{G}$	$\mathcal{G}; A$	$\mathcal{G}; A$	$\mathcal{G}; A$
$A; \mathcal{G}$	$A; \mathcal{G}$	$A; \mathcal{G}$	$A; \mathcal{G}$	$A; \mathcal{G}$	$A; \mathcal{G}$
$\mathcal{G}; A$	$\mathcal{G}; A$	$\mathcal{G}; A$	$\mathcal{G}; A$	$\mathcal{G}; A$	$\mathcal{G}; A$
$A \| \mathcal{G}$	$A \| \mathcal{G}$	$A \| \mathcal{G}$	$A \| \mathcal{G}$	$A \| \mathcal{G}$	$A \| \mathcal{G}$

Table 10.2 Parallel composition

$\parallel$	A	$\mathcal{G}$	$A;\mathcal{G}$	$\mathcal{G};A$	$A \parallel \mathcal{G}$
A	A	$A \parallel \mathcal{G}$	$A;\mathcal{G}$	$\mathcal{G};A$	$A \parallel \mathcal{G}$
$\mathcal{G}$	$A \parallel \mathcal{G}$	$\mathcal{G}$	$A \parallel \mathcal{G}$	$\mathcal{G};A$	$A \parallel \mathcal{G}$
$A;\mathcal{G}$	$A;\mathcal{G}$	$A \parallel \mathcal{G}$	$A;\mathcal{G}$	$A \parallel \mathcal{G}$	$A \parallel \mathcal{G}$
$\mathcal{G};A$	$A \parallel \mathcal{G}$	$\mathcal{G};A$	$A \parallel \mathcal{G}$	$\mathcal{G};A$	$A \parallel \mathcal{G}$
$A \parallel \mathcal{G}$	$A \parallel \mathcal{G}$	$A \parallel \mathcal{G}$	$A \parallel \mathcal{G}$	$A \parallel \mathcal{G}$	$A \parallel \mathcal{G}$

(1) $(A;\mathcal{G});(A \parallel \mathcal{G})$
$\equiv ((A;\mathcal{G}) \cap (A \cap \mathcal{G})) \cup (A;\mathcal{G})^{<}$
$\equiv (((A \cap \mathcal{G}) \cap A^{<}) \cap (A \cap \mathcal{G})) \cup (A^{<} \cup (A \cap \mathcal{G}^{<}))$
$\equiv ((A \cap \mathcal{G}) \cup (A^{<} \cap \mathcal{G})) \cup (A^{<} \cup (A \cap \mathcal{G}^{<}))$
$\equiv (A \cap \mathcal{G}) \cup (A^{<} \cup (A \cap \mathcal{G}^{<}))$
$\equiv (A \cap \mathcal{G}) \cup A^{<}$
$\equiv A;\mathcal{G}$

(2) $(A \parallel \mathcal{G});(A;\mathcal{G})$
$\equiv ((A \cap \mathcal{G}) \cap (A;\mathcal{G})) \cup (A \parallel \mathcal{G})^{<}$
$\equiv ((A \cap \mathcal{G}) \cap ((A \cap \mathcal{G}) \cup A^{<})) \cup ((A^{<} \cap \mathcal{G}) \cup (A \cap \mathcal{G}^{<}))$
$\equiv (A \cap \mathcal{G}) \cup (A^{<} \cap \mathcal{G}) \cup (A^{<} \cap \mathcal{G}) \cup (A \cap \mathcal{G}^{<})$
$\equiv (A \cap \mathcal{G}) \cup (A^{<} \cap \mathcal{G}) \cup (A \cap \mathcal{G}^{<})$
$\equiv A \cap \mathcal{G}$
$\equiv A \parallel \mathcal{G}$

(3) $(A;\mathcal{G}) \parallel (\mathcal{G};A)$
$\equiv ((A \cap \mathcal{G}) \cup A^{<}) \cap (\mathcal{G} \cap A) \cup P^{<})$
$\equiv (((A \cap \mathcal{G}) \cup A^{<}) \cap (\mathcal{G} \cap A)) \cup (((A \cap \mathcal{G}) \cup A^{<}) \cap P^{<})$
$\equiv ((A \cap \mathcal{G} \cup (A^{<} \cap P)) \cup ((A \cap P^{<}) \cup (A^{<} \cap P^{<}))$
$\equiv (A \cap \mathcal{G}) \cup (A \cap P^{<})$
$\equiv A \cap \mathcal{G}$
$\equiv A \parallel \mathcal{G}.$

Using these tables, a simple induction shows that all basic graph updates as defined above fall in the listed finite set.

Our final observation is that all these operations indeed induce *PDL*-definable betterness transformers, by the analysis in Section 10.5.1. □

We can even see a little more, by inspecting the form of the *PDL* programs for basic graph updates. These are all "flat forms" in the sense of Section 4.3, without iterated occurrences of the input relation R.[9]

Finally, we can extend the scope of the theorem a bit by allowing the "empty priority graph" that puts no restrictions on orderings. It clearly induces the *universal*

[9] This can also be used in an alternative proof for Theorem 10.22: In particular, the flat-format definable relations are closed under taking intersections.

relation between worlds, matching the universal modality, and this was indeed one more atomic relation in the *PDL*-format of Chapter 4.

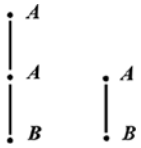
10.5.3 Obstacles to a Complete Match

Despite our positive result, we doubt that a general reduction is possible between world-level and graph-level preference transformations. For a start, this has to do with general obstacles to defining induced maps:

Fact 10.24 *The top deletion operation $del(\mathcal{G})$ is not PDL-definable.*

Proof Assume that some relation transformer σ defines graph top deletion in the sense of Definition 10.3.1. Consider the following two graphs (Fig. 10.2):

Fig. 10.2 Deletion



Here, $\mathcal{G}$ consists of two propositions A and B, A on top of B, while $\mathcal{G}'$ has three propositions, two copies of A on top of B, one A on top of the other. Then, syntactically, $del(\mathcal{G})$ is the new graph that contains only B, and while $del(\mathcal{G}')$ is the original graph $\mathcal{G}$.

Clearly, we then have the following two facts:

(a) $\leq_{\mathcal{G}} = \leq_{\mathcal{G}'}$, whereas (b) $\leq_{del(\mathcal{G})} \neq \leq_{del(\mathcal{G}')}$.

But, by our assumption, the relation transformer σ also satisfies

$\leq_{del(\mathcal{G})} = \sigma(\leq_{\mathcal{G}}) = \sigma(\leq_{\mathcal{G}'}) = \leq_{del(\mathcal{G}')}$.

This is the required contradiction. □

This counter-example illustrates a more general difference in dynamics: Deletion in graphs is hard to mimic at the level of induced orderings, as we throw away syntactic information that is not visible in the input model.[10]

But also conversely, there is no general match. Not all simply definable betterness transformers in Chapter 4 have inducing graph counterparts:

Fact 10.25 *Not all PDL-definable operations are graph-definable.*

[10] Similar difficulties with deletion were found for "agenda dynamics" in [83].

Proof Here is a counter-example. Not all betterness transformers preserve the base properties of reflexivity and transitivity. To see this, consider $?A; R$, that is: "keep the old relation only when A is true". This does not preserve reflexivity, as $\neg A$-worlds have no relations any more. So this relation-transformer cannot be defined using a partial priority graph. □

This betterness transformer amounts to a refusal to henceforth make betterness comparisons at worlds that lack property A. We regard such more general betterness transformers as bona fide mind changes of an agent.[11]

10.6 Discussion and Conclusion

The conclusion that we draw from our technical analysis in the preceding section is that, while priority dynamics and betterness dynamics have a significant overlap, they also have features of their own that resist reduction. This reinforces our earlier point, made for independent reasons of natural modeling, that we want both intrinsic and extrinsic preference as options for agents, with extrinsic preferences having their reasons encoded in priority graphs that are part of the model.

Two-level models Thus, we think that preference logic should work over "two-level models" having both priority structure and betterness relations (cf. [131]). Indeed, even in those cases where the betterness is indeed priority-based, and a "reduction" would be possible in principle, the two levels still offer an interesting option between two ways of describing dynamics. For instance, [48] present illustrations of how this freedom works as different strategies for achieving the same evaluation change for worlds in deontic and legal reasoning. But in general, two-level models may be "sui generis", having both reason-based extrinsic and primitive intrinsic betterness relations, and different actions at both levels changing these.

We will not explore the formal logic of this combined perspective in this chapter, but we end with a point of language dynamics.

Intrinsic preference, extrinsic preference, and language change In the spirit of this chapter, one can now use two-level models having both a graph-based extrinsic betterness relation and an intrinsic betterness relation that may reflect the agent's feelings or prejudices. But the contrast is relative, not absolute.[12] In particular, this contrast can be dissolved by means of a form of dynamics that is largely outside the scope of dynamic-epistemic logic, the main paradigm used in this chapter, viz. the phenomenon of *language change*.

---

[11] It is an interesting open problem if all *PDL*-flat-format-definable betterness transformers *that always generate pre-orders* are definable by syntactic graph updates.

[12] A nice illustration is deontic logic. If I obey the command of a higher moral authority, I may acquire an extrinsic preference, whose reason is obeying a superior. But for that higher agent, that same preference may be intrinsic: "The king's whim is my law".

An agent can literally *rationalize* a given intrinsic betterness relation by providing reasons for it. Sometimes, this may even be done within the given language, say, when the indifference classes ("clusters") of the relation are definable in terms of our modal formulas.[13] Alternatively, one can think of this someone else observing an agent's preferences, and postulating reasons for them. This is closer to the notion of "revealed preference" as studied in the economics literature: cf. [111] and our brief discussion in Chapter 12. But then, in general, the reasons may have to come from some other, usually richer language than the original one that we started with: We are witnessing a dynamic act of language creation.

Now, here is where our earlier representation result Theorem 10.5 comes in once more. What it shows is that pre-orders can always be rationalized in terms of suitable propositions partitioning the domain of worlds. And hence, it licenses, at least in principle, language extensions that can rationalize any given intrinsic betterness order. In our view this language dynamics is the right way of viewing the contrast between intrinsic and extrinsic preference. As a consequence, while not going back on our claim that two levels are needed in an account of preference, these levels of representation can be in flux, with changes from one to the other.[14]

Conclusion In this chapter, we have explored the idea that preference should be represented at two levels, both as betterness among worlds and as priority among propositions. We have shown that both levels support significant logical structure, and that the two are connected, though not reducible, in interesting ways. Our main results are technical, extending and explaining earlier work in preference dynamics. But we do think that the resulting picture of preference offers a more realistic view of how preferences are structured and can be changed, that can be applied to many areas. One of these is deontic reasoning, as we will demonstrate in Chapter 11 below. Another illustration is belief revision, as we have briefly indicated in Chapter 9.

This more richly structured picture of preference statics and dynamics is the final version of the theory proposed in this book. The remaining chapters will provide some exploration of its repercussions in other areas.

[13] It is an interesting technical problem just when such definitions are possible.

[14] This dynamic take on what may be called the "act of representation" has independent logical interest: Representation constructions suggest language dynamics.

Part VI
Applications and Discussions

Chapter 11
Deontic Reasoning

11.1 Introduction

Deontic logic is the logical study of normative concepts such as obligation, prohibition, permission and commitment. As we will see in this chapter, it is a very natural setting for with preference logic, both in its static versions (cf. [96, 183]) and in terms of the new dynamic systems of this book [48].

Normative concepts can be naturally made sense of in terms of an "ideality" ordering $\leq$ on possible worlds, as stated in Moore's work *Principia Ethica*:

> [...] to assert that a certain line of conduct is [...] absolutely right or obligatory, is obviously to assert that more good or less evil will exist in the world, if it is adopted, than if anything else be done instead.[1]

Depending on the properties of $\leq$, different logics will then be obtained. In particular, [96] starts with a $\leq$ which is only reflexive, moving then to total pre-orders.

In this chapter we are going to re-explore established ideas developed for the preference-based semantics of deontic logic in the light of recent studies in the modal logic of preference (in particular, the preceding parts of this book as well as [83]). Our treatment starts from the well-known semantics for dyadic obligation first introduced in [96], where dyadic obligations of the type "it is obligatory that φ under condition ψ" are interpreted by making use of a comparative "ideality relation" and the notion of maximality:

(1) $\mathfrak{M}, s \models O(\varphi \mid \psi)$ iff $Max(||\psi||_{\mathfrak{M}}) \subseteq ||\varphi||_{\mathfrak{M}}$.

where $||.||_{\mathfrak{M}}$ denotes the truth-set function of $\mathfrak{M}$ and $\mathfrak{M}$ is a model built on a possible worlds frame $\mathcal{F} = (S, \leq)$. In this frame, the states in S are ordered according to the ideality relation $\leq$. Although Formula (1) has been an object of many criticisms,[2] variations gave rise to several studies in the preference-based semantics of

[1] Cited in [78, p. 6].

[2] One criticism is that Formula (1) makes conditional obligations lack the property of antecedent strengthening (see [181]). This, however, makes perfect sense in our view as it is precisely what needs to follow from the idea of "most ideal worlds".

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_11, © Springer Science+Business Media B.V. 2011

deontic logic, which enjoyed considerable attention up till the 1990s.[3] In this light, the present chapter provides a "fresh look at an old idea" and offers an attempt at reviving this tradition by reinterpreting it in the light of recent developments in the modal logic of preference. More specifically, the insights given by Formula (1), will be extended by developing the following points: (i) the ideality order–which is much like our earlier "betterness" relation–can be fruitfully viewed as generated by a set of explicitly prioritized criteria; (ii) both the betterness relation and the priorities on criteria naturally support a dynamic point of view, giving rise to a richer system of deontic dynamics as based on preference logic. This latter point also brings deontic logic closer to the family of dynamic logics of belief and preference change [32].

More in detail, this chapter is structured as follows. Section 11.2 introduces the machinery of priority sequences from Chapter 7 to provide an original account of "contrary-to-duty-obligations (*CTDs*)". We will make a brief logical excursion showing how the deontically important notion of "best" can be defined in our framework. Section 11.3 then capitalizes on this richer modeling, and applies the type of perspective on deontics obtained by juxtaposing the "semantic" view of deontics yielded by betterness relations with the "syntactic" view of deontics yielded by priority sequences. In particular, we discuss two classical topics: Anderson's reduction, and the Chisholm Paradox. Section 11.4 continues with applying some further techniques introduced in Chapter 10 to study betterness dynamics, identifying, in particular, a correspondence between "syntactic" normative changes, and "semantic" ones at the level of the betterness relation. Section 11.5 concludes with our broader views. Finally, we have added an Appendix (based on the working paper [119]) showing how the ideas of this chapter find a natural continuation in a more linguistic topic related to deontic reasoning, viz. the semantics of imperative expressions.

11.2 Priorities, Betterness and CTDs

The betterness relation between states involved in the preference-based semantics of obligations (Formula 1) is, often, a sort of betterness derived from some kind of explicit "coding" of what is better in terms of relevant properties, as the following quote from St. Paul illustrates in a lively manner:

> It is good for a man not to touch a woman. But if they cannot contain, let them marry: for it is better to marry than to burn. [177, Ch. 7][4]

This is, in the terminology of deontic logic, a typical contrary-to-duty structure [153] expressing what states are best, what states are best among the non-best ones, and so on, up to a finite depth. In the following sections we will briefly discuss this type of structures in the light of notions and results developed in Part III and IV, and illustrate them by formalizing a classical example of *CTD* obligations.

---

[3] See [183] for an overview of this area of investigation.

[4] This passage is cited in [78, p. 6].

11.2.1 Priority Sequences

This section introduces the formal notions shaping the sort of ideality ordering with which we will work. To adopt the current context, we re-state some definitions in our new notations.

Definition 11.1 (P-sequence) Let $\mathcal{L}(\mathbf{P})$ be a propositional language built on the set of atoms $\mathbf{P}$, S a non-empty set of states and $\mathcal{I} : \mathbf{P} \longrightarrow 2^S$ a valuation function. A P-sequence for $\mathcal{I}$ is a tuple $\mathcal{B}^{\mathcal{I}} = (B, <)$ where[5]:

- $B \subset \mathcal{L}(\mathbf{P})$ with $|B| < \omega$;
- $<$ is a strict linear order on B;
- for all $\varphi, \psi \in B, \varphi < \psi$ iff $||\psi||_{\mathcal{I}} \subset ||\varphi||_{\mathcal{I}}$.

where $||\varphi||_{\mathcal{I}}$ denotes the truth-set of φ according to $\mathcal{I}$. The set of all P-sequences for $\mathcal{I}$ is denoted $\mathcal{B}^{\mathcal{I}}$. Given a P-sequence $\mathcal{B}^{\mathcal{I}}$ for $\mathcal{I}$ denote with $Max(\mathcal{B}^{\mathcal{I}})$ the maximum element of $\mathcal{B}^{\mathcal{I}}$. Also, we denote with $Max^{+}(\mathcal{B}^{\mathcal{I}})$ the maximum element of $\mathcal{B}^{\mathcal{I}}$ which has a non-empty denotation according to $\mathcal{I}$, if it exists, or $\top$ otherwise.

In other words, a priority sequence is a finite chain of distinct propositional formulae $\varphi_n < \cdots < \varphi_1$ from a language $\mathcal{L}(\mathbf{P})$ whose denotations form a finite ascending chain of sets $||\varphi_1||_{\mathcal{I}} \subset \cdots \subset ||\varphi_n||_{\mathcal{I}}$.[6] For this reason they can be referred to as lists $\varphi_1, \ldots, \varphi_n$ of elements with $|B| = n$.[7] It is worth stressing that a P-sequence is always relative to a valuation function for its propositions. If $||\varphi_1||_{\mathcal{I}}$ and $||\varphi_2||_{\mathcal{I}}$ are incomparable with respect to set-theoretic inclusion, then they cannot be part of the same P-sequence for $\mathcal{I}$.

We now define a simple way to order states according to a given P-sequence.

Definition 11.2 (deriving preferences from P-sequences) Let $\mathcal{B} = (B, <)$ be a P-sequence, S a non-empty set of states and $\mathcal{I} : \mathbf{P} \longrightarrow 2^S$ a valuation function. The preference relation $\leq_{\mathcal{B}}^{IM} \subseteq S^2$ is defined as follows:

(2) $s \leq_{\mathcal{B}}^{IM} s' := \forall \varphi \in B : s \in ||\varphi|| \Rightarrow s' \in ||\varphi||$.

where IM is just a mnemonics for "implication". Given a P-sequence $\mathcal{B}$ for a valuation $\mathcal{I}$, Formula (2) generates also a Kripke model $\mathfrak{M}_{\mathcal{B}}^{IM} = (S, \leq_{\mathcal{B}}^{IM}, \mathcal{I})$.

Intuitively, Definition 11.2 orders states in S according to which elements of the P-sequence they satisfy. If a state satisfies a property in the sequence, then it also satisfies, by Definition 11.1, all $<$-worse properties in the sequence. We therefore obtain an order on states with the following properties.

---

[5] When no confusion arises we will often drop the superscript in $\mathcal{B}^{\mathcal{I}}$.

[6] When no confusion arises we will often drop the subscript in $||\varphi||_{\mathcal{I}}$.

[7] It might be instructive to notice that Definition 11.1 could be restated by requiring the elements of the sequence to be disjoint in $\mathcal{I}$, instead of being ordered according to a finite $\subseteq$-chain (see [131] for further details).

Fact 11.3 (Properties of $\leq_{\mathcal{B}}^{IM}$) *Let $\mathcal{B} = (B, <) = (\varphi_1, \ldots, \varphi_n)$ be a P-sequence for $\mathcal{I} : \mathbf{P} \longrightarrow 2^S$. It holds that:*

(1) *Relation $\leq_{\mathcal{B}}^{IM}$ is a total pre-order*
(2) *If $\varphi_i < \varphi_j$ then for all $s \in ||\varphi_i||, s' \in ||\varphi_j||$: $s \leq_{\mathcal{B}}^{IM} s'$;*
(3) *If $\varphi_i < \varphi_j$ then for all $s \in ||\varphi_i \wedge \neg\varphi_j||, s' \in ||\varphi_j||$: $s <_{\mathcal{B}}^{IM} s'$.*

Proof The first claim is straightforward as to reflexivity, transitivity and connectedness. As to converse well-foundedness, note that P-sequences are finite. So, if the cardinality of a P-sequence is n, it generates a total pre-order consisting of at most $n + 1$ clusters of equally good states. Hence the strict part of the pre-order contains only bounded chains. The second and third claims follow directly from Definitions 11.1 and 11.2. □

The following is worth noticing. Given a P-sequence $\varphi_n < \cdots < \varphi_1$, the $\leq_{\mathcal{B}}^{IM}$-minimal states in S are the one satisfying $\neg\varphi_n$, if such states exists. In fact, it can be the case that $||\varphi_n|| = S$, i.e., $\varphi_n = \top$. In such a case, the $\leq_{\mathcal{B}}^{IM}$-minimal states are therefore the states satisfying $\neg\varphi_{n-1}$. This suggests that any $\varphi_n < \cdots < \varphi_1$ P-sequence such that $||\varphi_n|| \neq S$ could be completed to a sequence $\varphi_{n+1} < \varphi_n < \cdots < \varphi_1$ where $\varphi_{n+1} = \top$.

11.2.2 P-Sequences and CTDs

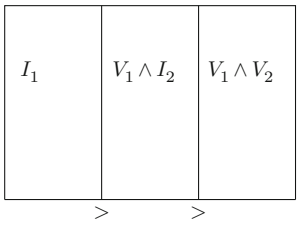
The example below is a "classic" of deontic logic and illustrates in a straightforward way the problem of *CTDs* [77]. We show how P-sequences can represent it in a natural way.

Example 11.4 (gentle murder) Here is the example:

> Here is the problem: Let us suppose a legal system which forbids all kinds of murder, but which considers murdering violently to be a worse crime than murdering gently. [...] The system then captures its views about murder by means of a number of rules, including these two:
>
> 1. It is obligatory under the law that Smith not murder Jones.
> 2. It is obligatory that, if Smith murders Jones, Smith murders Jones gently. [77, p. 194]

The scenario makes explicit two classes of ideality: a class (let us call it I_1) in which Smith does not murder Jones, i.e. $\mathsf{I}_1 := \neg m$; another one (let us call it I_2), in which either Smith does not murder Jones or he murders him gently, i.e., $\mathsf{I}_2 := \neg m \vee (m \wedge g)$. We thus have a P-sequence $\mathcal{B}$ such that $\mathsf{I}_2 < \mathsf{I}_1$. Such P-sequence is sufficient to order the states – according to the corresponding $\leq_{\mathcal{B}}^{IM}$ relation – in three clusters such that the most ideal states are the ones satisfying I_1, the worse but not worst states are the ones that satisfy $\mathsf{V}_1 := \neg\mathsf{I}_1$ but at the same time I_2 and, finally, the worst states are the ones satisfying $\mathsf{V}_2 := \neg\mathsf{I}_2$ (and hence V_1 too). This is depicted in Fig. 11.1.

Fig. 11.1 Gentle murder



To sum up, the intuition behind a P-sequence $p_1, \ldots, p_n$ for a given interpretation function is that each atom p_i gives rise to a bipartition $\{\mathcal{I}(p_i), -\mathcal{I}(p_i)\}$ of the domain of discourse S, and the more we move towards the right-hand side (i.e., the bottom) of the sequence the more atoms p_i are falsified. As shown by Example 11.2, in a deontic reading this simply means that, the more we move towards the right-hand side of the sequence the more violations hold.

11.2.3 "To Make the Best of Sad Circumstances"

Although Example 11.2 has nicely illustrated how a *CTD* structure can be rendered by a P-sequence, it still remains to be shown how the basic *CTD* reasoning operates on such structures. The idea is to express how "to make the best out of sad circumstances" [96]. On a P-sequence this means, intuitively, take the best states that survive the "sad circumstances". To make this precise we have to introduce the following refinement of Definition 11.1, which relativizes the notion of P-sequence to the occurrence of given circumstances.

Definition 11.5 (restricted P-sequences) Let $\mathcal{B} = (B, <)$, with $|B| = n$ and $B \subset \mathcal{L}(\mathbf{P})$, be a P-sequence for $\mathcal{I} : \mathbf{P} \longrightarrow 2^S$. The restriction of $\mathcal{B}$ to a formula ψ of $\mathcal{L}(\mathbf{P})$ is a structure $\mathcal{B}^\psi = (B^\psi, <^\psi)$ where:

- $B^\psi := \{\varphi_i \wedge \psi \mid \varphi_i \in B\}$;
- $<^\psi := \{(\varphi_i \wedge \psi, \varphi_j \wedge \psi) \mid (\varphi_i, \varphi_j) \in <\}$.

Given a restricted P-sequence $\mathcal{B}^\psi$ for $\mathcal{I}$, we denote with $Max(\mathcal{B}^\psi)$ the maximum element of $\mathcal{B}^\psi$. Also, we denote with $Max^+(\mathcal{B}^\psi)$ the maximum element of $\mathcal{B}^\psi$ which has a non-empty denotation according to $\mathcal{I}$, if it exists, or ψ otherwise.

Intuitively, the restriction of a P-sequence with respect to (the interpretation) of a formula ψ simply intersects the elements of the original P-sequence with ψ and keeps the original linear order. The result of such operation bears effects for the Max^+ of the P-sequence. Typically, $Max^+(\mathcal{B})$ might differ from $Max^+(\mathcal{B}^\psi)$ as $||1_\mathcal{B} \wedge \psi||$ (i.e., the intersection of the maximum element of $\mathcal{B}$ with ψ) might be empty. Notice that if all elements in $\mathcal{B}^\psi$ turn out to be empty for a given $\mathcal{I}$, $Max^+(\mathcal{B}^\psi)$ is taken to be ψ itself (cf. Definition 11.1).

Example 11.6 (gentle murder (continued)) Consider the P-sequence of the Gentle murder given in Example 11.2: $\mathcal{B} = (1_1, 1_2)$. The restricted P-sequence $\mathcal{B}^{V_1}$ is

then $(\mathsf{I}_1 \wedge \mathsf{V}_1, \mathsf{I}_2 \wedge \mathsf{V}_1)$. In such a sequence the top element has necessarily an empty denotation. This means that the best among the still available states are the states $Max^+(\mathcal{B}^{\mathsf{V}_1}) = \mathsf{I}_2 \wedge \mathsf{V}_1$ (see Definition 11.5). Another interesting restricted P-sequence in the Gentle murder context is $\mathcal{B}^{\mathsf{V}_2}$, which describes what the original P-sequence prescribes under the assumption that also the *CTD* obligation "kill gently" has been violated. In this case $Max^+(\mathcal{B}^{\mathsf{V}_2}) = \mathsf{V}_2$, that is to say, if also the last *CTD* obligation has been violated, then we end up in a set of all equally bad states. This illustrates a characteristic feature of all finite *CTD* structures.

Stated more positively, our approach, including the logical elaboration to follow, provides a simple perspective the robustness of norms and laws viewed as *CTD* structures: They can still function when transgressions have taken place.

Other major deontic puzzles can be dealt with as in Examples 11.2–11.6. As an illustration, Section 11.3.2 will propose an analysis of the Chisholm Paradox. Before moving to the next section, it is worth mentioning that representing *CTD* structures as finite chains of properties is not a new idea in the literature on deontic logic. To the best of our knowledge, this idea was first adumbrated informally in [78]. The first formal account of *CTD* structures as sequences of formulae is to be found in [86], where an elegant Gentzen calculus is developed for handling formulae of the type $\varphi_1 \otimes \ldots \otimes \varphi_n$ with $\otimes$ a connective representing a sort of "sub-ideality" relation in a *CTD* structure. Unlike this proof-theoretic approach, our approach is geared towards semantics and aims at connecting such *CTD* structures to modal logics interpreted on orders, and ultimately to conditional obligations in the standard maximality-based semantics (Chapter 3).[8]

11.2.4 'Best' in Modal Betterness Logic

In this chapter we assume all results about modal betterness logic obtained in Chapter 3. Here, we repeat some relevant parts of our discussion of the expressive power of that language, and see how to define "best", as this notion plays a central role in deontic contexts.

The very first semantics for dyadic deontic logic [96] interpreted formulae $O(\varphi \mid \psi)$ as "all the best ψ-states are φ" (Formula 1). Within our logic language, a maximality operator can be defined as follows:

(3) $[\mathsf{Best}(\psi)]\,\varphi := U(\psi \rightarrow \langle \leq \rangle(\psi \wedge [\leq](\psi \rightarrow \varphi)))$.

That is, the best ψ-states are φ if and only if, for all states, either they are not ψ or there is a better ψ-state such that all states above it are either not ψ or φ. As we

[8] With respect to this, the interesting question arises of whether the Gentzen calculus developed in [86] is complete for our semantics, or whether it could be embedded in the logic presented in Chapter 3.

mentioned before, this definition was first proposed, together with some variants, by [53] and [55]. Here, we restated it using the universal modality.[9]

Similar to the case of conditional preference and conditional beliefs, we should assume converse well-foundedness for the maximality operator, to avoid empty sets of maximals for some formulae.

Example 11.7 (gentle murder (continued)) Consider the P-sequence for valuation $\mathcal{I}$ of the Gentle murder introduced in Example 11.2, and let $\leq_{\mathcal{B}}^{IM}$ be the total pre-order generated by that sequence. We have that, for any state s in the model $\mathfrak{M}_{\mathcal{B}}^{IM} = (S, \leq_{\mathcal{B}}^{IM}, \mathcal{I})$:

$$\mathfrak{M}_{\mathcal{B}}^{IM}, s \models [\mathsf{Best}(\top)]\, \mathsf{I}_1.$$
$$\mathfrak{M}_{\mathcal{B}}^{IM}, s \models [\mathsf{Best}(\mathsf{V}_1)]\, \mathsf{I}_2.$$
$$\mathfrak{M}_{\mathcal{B}}^{IM}, s \models [\mathsf{Best}(\mathsf{V}_2)]\, \mathsf{V}_2.$$

It is easy to see that the maximality statements formulated above correspond with the analysis via restricted P-sequences provided in Example 11.6. In a way, Examples 11.6 and 11.7 depict two complementary views of obligation. The next section investigates this connection in more detail.

11.3 Deontics as Founded on Classification and Betterness

We briefly sum up the technical results thus far about our logical framework:

- *CTD* structures can be represented syntactically as P-sequences;
- P-sequences determine total pre-orders with a conversely well-founded strict part, and we can reason about such structures within a suitable modal logic.

Given a P-sequence $\mathcal{B}$, these two insights suggest two ways of defining dyadic obligation operators: (i) what ought to be the case is what the best non-empty property of the P-sequence logically implies; (ii) what ought to be the case is what holds in the best states. In formulae:

(4) $O(\varphi \mid \psi) := U(Max^{+}(\mathcal{B}^{\psi}) \rightarrow \varphi)$.
(5) $O(\varphi \mid \psi) := [\mathsf{Best}(\psi)]\, \varphi$.

where we recall that $Max^{+}(\mathcal{B}^{\psi})$ denotes the best non-empty element of $\mathcal{B}$ in the restriction of $\mathcal{B}$ to ψ (see Definition 11.5). The second formula is the very first semantics of dyadic deontic logic, already discussed in Chapter 3. The first is reminiscent of the Andersonian-Kangerian reduction of deontic logic.[10]

---

[9] Other variants of Formula (3) are possible. For instance, [83, Ch. 3] proposes $[\mathsf{Best}(\psi)]\, \varphi := U((\psi \wedge \neg\langle < \rangle \psi) \rightarrow \varphi)$ for the case that models are finite.

[10] We will come back to this aspect in Section 11.3.1.

Formulae (4) and (5) are an interesting dichotomy in the definition of "ought". While Formula (4) resorts to a "classification" stating that φ is what necessarily follows from a given (syntactic) standard of behavior – the P-sequence – under circumstances φ, Formula (5) simply resorts to a given (semantic) betterness relation on states and the notion of maximality. The present section shows how they coincide, thus illustrating the two faces of the notion of obligation. To say it with St. Thomas Aquinas:

> Voluntas [...] bonorum consonat legi, a qua malorum voluntas discordat. [178, ia.2ae 96,5]

That is to say, the preferences of the good men are in accordance with the law, and those of the bad men in discordance.

11.3.1 Connecting Obligations to What is the Best

Given a P-sequence, there is full accordance between defining what is obligatory as what is true in the best states of the order yielded by the P-sequence, or as what logically follows from the highest ranked property in the P-sequence. In other words, there is full correspondence between the "letter" of the law (the P-sequence), and its content (the betterness ordering).

Theorem 11.8 ("What is obligatory is that which follows from the best") *Let $\mathcal{B} = (B, <)$ be a P-sequence. For any model $\mathfrak{M}^{IM}_{\mathcal{B}}$ derived from $\mathcal{B}$ as in Definition 11.2 and state s it holds that:*

$$\mathfrak{M}^{IM}_{\mathcal{B}}, s \models [\mathsf{Best}(\psi)]\,\varphi \iff \mathfrak{M}^{IM}_{\mathcal{B}}, s \models U(Max^{+}(\mathcal{B}^{\psi}) \rightarrow \varphi).$$

where $\mathcal{B}^{\psi}$ is the restriction of $\mathcal{B}$ to ψ (Definition 11.5).

Proof A proof can be obtained by the subsequent application, in this order, Definition on truth conditions, Definitions 11.2 and 11.5:

$$\begin{aligned}
\mathfrak{M}^{IM}_{\mathcal{B}}, s \models [\mathsf{Best}(\psi)]\,\varphi &\iff \forall s' \in Max((\|\psi\|, \leq^{IM}_{\mathcal{B}})) : \mathfrak{M}^{IM}_{\mathcal{B}}, s' \models \varphi \\
&\iff \forall s' \in \|\psi\| \text{ s.t. } [\forall s'' \in \|\psi\| : s'' \leq^{IM}_{\mathcal{B}} s'] : \mathfrak{M}^{IM}_{\mathcal{B}}, s' \models \varphi \\
&\iff \forall s' \in \|\psi\| \text{ s.t. } [\forall s'' \in \|\psi\|, \forall \mathsf{l}_i \in B : s'' \in \mathcal{I}(\mathsf{l}_i) \Rightarrow \\
&\qquad s' \in \mathcal{I}(\mathsf{l}_i)] : \mathfrak{M}^{IM}_{\mathcal{B}}, s' \models \varphi \\
&\iff \forall s' \in \|Max^{+}(\mathcal{B}^{\psi})\| : \mathfrak{M}^{IM}_{\mathcal{B}}, s' \models \varphi \\
&\iff \forall s' \text{ s.t. } \mathfrak{M}^{IM}_{\mathcal{B}} \models Max^{+}(\mathcal{B}^{\psi}) : \mathfrak{M}^{IM}_{\mathcal{B}}, s' \models \varphi \\
&\iff \forall s' : \mathfrak{M}^{IM}_{\mathcal{B}}, s' \models Max^{+}(\mathcal{B}^{\psi}) \rightarrow \varphi \\
&\iff \mathfrak{M}^{IM}_{\mathcal{B}}, s \models U(Max^{+}(\mathcal{B}^{\psi}) \rightarrow \varphi)
\end{aligned}$$

□

In words, φ is the best that can hold given ψ if and only if φ is what is required by ideality under the circumstances ψ. Or, to put it yet otherwise, φ is what the law

says is a primary obligation under circumstances ψ. We might suggestively say that Theorem 11.8 captures the ethical truism that "what ought to be the case is what is best under the given circumstances".

11.3.1.1 Anderson's Reduction and P-Sequences

Anderson's [5] and Kanger's [120] reduction of deontic logic consists in a definition of O-formulae to alethic modal logic $\Box$-formulae containing a designated violation or ideality constant:

(6) $O\varphi :=\Box(\neg\varphi \rightarrow \mathsf{V})$.
(7) $O\varphi :=\Box(\mathsf{I} \rightarrow \varphi)$.

These are obviously equivalent under the assumption that $\mathsf{V} \leftrightarrow \neg\mathsf{I}$ is a valid principle. It is well-known that this reductionist view of deontic logic inherits a number of the weaknesses of deontic logic – among which the impossibility of representing *CTDs* satisfactorily. In this section, we briefly show how P-sequences offer a natural extension to Anderson's and Kanger's reduction.

We call a Kangerian-Andersonian P-sequence any P-sequence consisting of ideality atoms (or an inverse sequence of violation atoms). We have the following result. The following corollary further illustrates Theorem 11.8 in the light of Kangerian-Andersonian P-sequences.

Corollary 11.9 (obligations from better to worse) *Let $\mathcal{B} = (B, <) = (\mathsf{I}_1, \ldots, \mathsf{I}_n)$ be a Kangerian-Andersonian P-sequence for $\mathcal{I}$ of n non-empty elements. For any model $\mathfrak{M}^{IM}_{\mathcal{B}}$ and state s it holds that:*

(8) $\mathfrak{M}^{IM}_{\mathcal{B}}, s \models O(\varphi \mid \top) \Longleftrightarrow \mathfrak{M}^{IM}_{\mathcal{B}}, s \models U(\mathsf{I}_1 \rightarrow \varphi)$.
(9) $\mathfrak{M}^{IM}_{\mathcal{B}}, s \models O(\varphi \mid \mathsf{I}_1) \Longleftrightarrow \mathfrak{M}^{IM}_{\mathcal{B}}, s \models U(\mathsf{I}_1 \rightarrow \varphi)$.
(10) $\mathfrak{M}^{IM}_{\mathcal{B}}, s \models O(\varphi \mid \mathsf{V}_i) \Longleftrightarrow \mathfrak{M}^{IM}_{\mathcal{B}}, s \models U(\mathsf{I}_{i+1} \rightarrow \varphi)$ *for* $1 \leq i < n$.
(11) $\mathfrak{M}^{IM}_{\mathcal{B}}, s \models O(\varphi \mid \mathsf{V}_n) \Longleftrightarrow \mathfrak{M}^{IM}_{\mathcal{B}}, s \models U(\mathsf{V}_n \rightarrow \varphi)$.

where $O(\psi \mid \varphi)$ is defined by Formula (4) or (5).

Proof Formulae (9)–(10) are instances of Theorem 11.8. □

Formula (8) establishes that an unconditional obligation $O(\varphi \mid \top)$ corresponds to what the most ideal states dictate. The corollary also shows how the content of obligations changes as we move from most ideal to least ideal circumstances. If we are in most ideal states, where I_1 holds, than what ought to be the case is in fact what already is the case (Formula (9)). Formula (10) states that, if it is the case that the ith ideality has been violated, where the ith is not the last one in the sequence, then what ought to be is what follows from the $(i + 1)$th ideality. In other words, if a norm is violated we look at the one telling us what to do if that is the case, that is, we move downwards in the P-sequence. Finally, if we are in least ideal states,

where I_n is false, then what ought to be the case is again what is already the case (Formula (11)).

All in all, the corollary offers a reinterpretation and generalization of the Andersonian-Kangerian reduction, where statements of the type $\Box(\mathsf{I}_i \rightarrow \varphi)$ are interpreted as assertions concerning the relative inclusions of state labels (e.g., "all I_i-states are φ-states"), or of the ideality and sub-ideality relations proposed in [116], where such relations are taken to be the total pre-orders generated by a P-sequence.

11.3.2 Chisholm Paradox Revisited

Next, we put our ideas to work on another classic of deontic reasoning, the Chisholm Paradox. We follow the presentation given in [7].

1. It ought to be that Smith refrains from robbing Jones.
2. Smith robs Jones.
3. If Smith robs Jones, he ought to be punished for robbery.
4. It ought to be the case that, if Smith refrains from robbing Jones, he is not punished for robbery.

Once "Smith robbing Jones" is represented by r and "Smith refraining from robbing Jones" by $\neg r$ and, similarly, "Smith being punished" by p while "Smith not being punished" by $\neg p$, this set of ordinary language sentences – also called the Chisholm's set – can receive the following four formalizations within a propositional modal language with modal operator O:

$$
\begin{array}{llll}
i)\ O\neg r & ii)\ O\neg r & iii)\ O\neg r & iv)\ O\neg r \\
\quad r & \quad r & \quad r & \quad r \\
\quad r \rightarrow Op & \quad O(r \rightarrow p) & \quad r \rightarrow Op & \quad O(r \rightarrow p) \\
\quad \neg r \rightarrow O\neg p & \quad O(\neg r \rightarrow \neg p) & \quad O(\neg r \rightarrow \neg p) & \quad \neg r \rightarrow O\neg p
\end{array}
$$

But it becomes quickly clear that none of such formalization work: In (i) the 4th statement $\neg r \rightarrow O\neg p$ is a logical consequence of the 2nd r, which is not the case in the ordinary language formulation; in (ii) the 3rd statement $O(r \rightarrow p)$ is a logical consequence of the 1st $O\neg r$, which is also not the case in the ordinary language formulation; finally, in (iii) and (iv) both Op and $O\neg p$ are logical consequences of the set and hence, by standard modal principles, $O\bot$ also follows, which should not be satisfiable O-formula.

The preceding Chisholm set of constraints displays the same structure as the earlier gentle murder scenario. So let us see how it can be dealt with in our framework. The first essential step is to have a model-theoretic look at the "paradox", just like we did in Example 11.2.

Example 11.10 (Chisholm's models) Let $\mathbf{P} := \{p, r, \mathsf{I}_1, \mathsf{I}_2, \mathsf{V}_1, \mathsf{V}_2\}$ and assume $\mathsf{V}_1 := \neg\mathsf{I}_1$ and $\mathsf{V}_2 := \neg\mathsf{I}_2$. Consider the following P-sequence for valuation $\mathcal{I}$:

$$\mathcal{B} = (\{\mathsf{I}_1, \mathsf{I}_2\}, \{(\mathsf{I}_2, \mathsf{I}_1), (\mathsf{I}_1, \mathsf{I}_1), (\mathsf{I}_2, \mathsf{I}_2)\})$$

Here type I_1 is strictly preferred to type I_2 and the resulting betterness model $\mathfrak{M}^{IM}_{\mathcal{B}} = (S, \leq^*_{\mathcal{B}}, \mathcal{I})$. A model $\mathfrak{M}^{IM}_{\mathcal{B}} = (S, \leq^{IM}_{\mathcal{B}}, \mathcal{I})$ is a Chisholm's model if the following formulae are valid:

(12) $O(\neg r \mid \top)$.
(13) $O(p \mid r)$.
(14) $O(\neg p \mid \neg r)$.

or, equivalently by Theorem 11.8, if the following formulae are valid:

(15) $U(\mathsf{I}_1 \rightarrow \neg r)$.
(16) $U(r \rightarrow (\mathsf{I}_2 \rightarrow p))$.
(17) $U(\mathsf{I}_1 \rightarrow \neg p)$.

Recall that the P-sequence for $\mathcal{I}$ requires $\mathcal{I}(\mathsf{V}_2) \subset \mathcal{I}(\mathsf{V}_1)$, and therefore that $U((\mathsf{I}_1 \rightarrow \mathsf{I}_2) \wedge \neg(\mathsf{I}_2 \rightarrow \mathsf{I}_1))$ is a validity. The Chisholm's scenario is thus naturally modeled by the r-states of a Chisholm's model, and in such states the $\leq^{IM}_{\mathcal{B}}$-maximal states are p-states.

In the above representations, no paradox arises. Instead, the formalization helps making explicit a semantically precise interpretation of the ordinary language formulation of the Chisholm's set. Formulae (12) and (15)–which are equivalent by Corollary 11.9–all state that the most ideal states are $\neg r$-states. Formulae (13) and (14) represent the *CTD* obligation of the Chisholm's scenario: Under the circumstance that r then it is most ideal that p. Finally, and probably most interestingly, Formulae (14) and (17) make clear that, in fact, the most ideal states are states in which no punishment occurs, since, by Corollary 11.9, it follows that they are all equivalent to $O(\neg p \mid \top)$.

To get back to ordinary language, Example 11.10 shows that a natural and consistent interpretation of the Chisholm's scenario would go as follows:

1. It is most ideal that Smith refrains to rob Jones;
2. Smith robs Jones;
3. The most ideal states under the assumption that Smith robs Jones are states in which Smith is punished;
4. It is most ideal that Smith is not punished.

To conclude, notice also that this reading of the scenario fits again the type of structure depicted earlier in Fig. 11.1, which was introduced originally to illustrate the gentle murder example.

11.4 Betterness Dynamics and Deontics

So far we have proposed a rich structured model for deontics under static circumstances. But deontic reasoning is crucially also about change, as events happen that change our evaluation of states-of-affairs in the world. Our concern in this section is *deontic dynamics*. In a way, we have already addressed a notion of dynamics by dealing with conditional obligations (Formula (1)), that is, the simplest type of deontic dynamics where changes in the condition determine changes in the normative consequences. This type of dynamics does not modify the betterness relation and is achieved via a maximality-based definition. In the remaining of this section we will address "genuine" changes in the betterness relation, as well as in the explicit codifications of betterness represented by the P-sequences.

11.4.1 Two Level Dynamics in Deontics

In the current framework we can handle dynamical changes that are located both at the level of P-sequences and at the level of states. The operations that were considered in Chapter 10 for those two kinds of dynamics apply naturally here.

Before getting started with the formal definitions, in order to illustrate the intuitions backing this section we add a dynamic twist to the "classic" example used in the presentation of the static framework: the gentle murder.

Example 11.11 (gentle murder dynamified) Let us start with the P-sequence consisting of $\mathcal{B} = (\neg m)$. By Definition 11.2, this generates a dychotomous total pre-order where all $\neg m$ states are above all m states: "It is obligatory under the law that Smith not murder Jones". Suppose this is the given deontic state-of-affairs. Now, how can we refine it in order to introduce the sub-ideal obligation to kill gently: "It is obligatory that, if Smith murders Jones, Smith murders Jones gently"? Or, in other words, how can we model the introduction, in the legal code, of this contrary-to-duty? Intuitively, this can happen in two ways:

1. We refine the given betterness ordering "on the go" by requesting a further bipartition of the violation states, putting the $m \wedge g$-states above the $m \wedge \neg g$-states. This can be seen as the successful execution of a command of the sort "if you murder then murder gently!".
2. We introduce a new law "from scratch", where $m \rightarrow g$ is explicitly stated as a class of possibly sub-ideal states. This can be seen as the enactment of a new P-sequence altogether: $(\neg m, m \rightarrow g)$, which is the P-sequence we have already encountered in Example 11.2.[11]

The example illustrates how a *CTD* sequence can be dynamically created either by the utterance of a sequence of commands each stating what ought to be the case in a sub-ideal situation, or by the direct enactment of a whole P-sequence. These two points of view on the creation of obligations of a *CTD* type are the dynamic

[11] Notice that $m \rightarrow g$ is equivalent to $\neg m \vee (m \wedge g)$.

counterpart of the two-level deontic logic studied in the preceding sections. In what follows our focus are the close ties between changes on betterness relation and changes at the level of P-sequences.

To formalize such connections, recall the technical notion in Definition 10.19 from Chapter 10. It states that a priority change α induces a betterness change σ if and only if the new total pre-order obtained via σ is the same as the one derived from the P-sequence after the changes dictated by α. More precisely, given a definition for deriving betterness relations from P-sequence (e.g., Definition 11.2), for any P-sequence $\mathcal{B}$ and new formula φ, the following holds:

$$\sigma(\leq_{\mathcal{B}}, \varphi) = \leq_{\alpha(\mathcal{B},\varphi)} .$$

Now, let σ be the betterness upgrade operation defined in Definition 4.10, and let α be the operation of postfixing a P-sequence $\mathcal{B}$ with a new formula φ (in symbols $\mathcal{B}; \varphi$), that is, of adding a least propositional formula in the P-sequence. In Chapter 10, we saw how the latter operation induces the former under Definition 11.2.

Theorem 11.12 (correspondence of the two-level dynamics in CTDs) *Let $\leq$ be derived from a P-sequence $\mathcal{B} = (B, <) = (\mathsf{l}_1, \ldots, \mathsf{l}_n)$, and let φ be of the form $\neg\mathsf{l}_n \rightarrow \psi$, with ψ a propositional formula such that $||\psi|| \supset \emptyset$. The following diagram then commutes:*

$$\begin{array}{ccc} \mathcal{B} & \xrightarrow{:\varphi} & \mathcal{B}; \varphi \\ {\scriptstyle IM}\downarrow & & \downarrow{\scriptstyle IM} \\ \left(S, \leq_{\mathcal{B}}^{IM}\right) & \xrightarrow{\Uparrow\varphi} & \left(S, \Uparrow \varphi\left(\leq_{\mathcal{B}}^{IM}\right)\right) \end{array}$$

This theorem is, in a way, a dynamic variant of the kind of static correspondence shown in Theorem 10.21 in Chapter 10. It makes the two faces of deontics explicit, also from a dynamic point of view. It is then easy to see that Example 11.11 represents precisely an instance of Theorem 11.12 where $\mathsf{l} := \neg m$: In words, the same

$$\begin{array}{ccc} (\neg m) & \xrightarrow{:m\rightarrow g} & (\neg m); m \rightarrow g \\ {\scriptstyle IM}\downarrow & & \downarrow{\scriptstyle IM} \\ \left(S, \leq_{(\neg m\rangle}^{IM}\right) & \xrightarrow{\Uparrow m\rightarrow g} & \left(S, \Uparrow m \rightarrow g\left(\leq_{(\neg m)}^{IM}\right)\right) \end{array}$$

deontic change can be obtained both by "refining" the order dictated by the given law, via subsequent orders (of a certain syntactic form), as well as by enacting a new "law" which correspondingly extends the given one.

Although this representation of Example 11.11 nicely captures a specific type of dynamics involved in *CTD*s, it is clearly just one example of the possible applications of a two-level analysis of deontic scenarios. For instance, at the level of P-sequences, besides adding a new proposition, we can also delete criteria, and study

the effect of that as a transformation on the betterness ordering. Also, so far we have only considered just one way (i.e., Definition 11.2) of deriving deontic betterness from a P-sequence. Will other definitions taking a norm system to individual preferences still allow for the commutative diagrams of Theorem 11.12? All these are options for future investigation opened up by the present framework.

11.4.2 Discussion: Betterness Dynamics and Norm Change

The dynamic aspects of norms – the so-called *norm change* problem – have recently gained much attention from researchers in deontic logic, legal theory and multi-agent systems. Before concluding this section on deontic dynamics we briefly want to put our work in perspective with some of the more recent contributions available in the literature on this topic, highlight similarities and future research lines.

In our view, existing approaches to norm change fall into two main groups. In syntactic approaches – inspired by legal practice – norm change is an operation performed directly on the explicit provisions in the "code" of the normative system, [51, 87, 88]. In semantic approaches, norm change follows the dynamic logic update paradigm (e.g., [12]). Our betterness dynamics belongs to this latter group. Thus, it can be naturally related to the sort of context dynamics – and the related dynamics of counts-as rules – studied in [12] via the bridge offered by Corollary 11.9: Obligations defined via ideality and maximality are special kinds of classifications of an Andersonian-Kangerian type.

Also, the dynamic logic connection enables a unified treatment of *two kinds of change* that mix harmoniously in deontic reasoning: "information change" given a fixed normative order, and "evaluation change" when the normative order itself changes. Their interplay reflects the *entanglement* of obligation, knowledge and belief studied in Chapters 5, 8, and [149].

Finally, despite its semantic flavor, the study of dynamics we propose here bridges very well with more syntactic analysis of norm change. In fact, Theorem 11.12 can be viewed as establishing a precise match between changes at the level of models with changes at the level of syntax of a normative code, i.e., the P-sequences. And this is not an accident of our simple total orders. Although our analysis has focused on linearly ordered P-sequences and the induced total pre-orders – fitting for *CTD*s – the logical machinery presented here easily generalizes to pre-orders (see [6] and Chapters 7, 10). Future work along these lines would look at more elaborate syntactic representations in line with our approach. These might include graph structures of criteria and laws, maintaining a normative syntax-semantics correspondence along the lines of Theorem 11.12 (see a recent study of [40]).

11.5 Conclusion

This chapter has revisited the preference-based semantics of deontic logic first presented in [96], and then developed it further into a richer medium for analyzing

deontic reasoning. This was done in two related ways: by introducing a two-level perspective on deontic ideality which enables a two-faceted modal analysis of deontic concepts (Theorem 11.8), and also a rich view of deontic dynamics as betterness change in tandem with norm change (Theorem 11.12).

Although our analysis has focused on linearly ordered P-sequences and their generated total pre-orders–particularly fitting for $CTDs$–the logical machinery presented here easily generalizes to the weaker structures of pre-orders (see [6] and Chapter 10). A generalization to arbitrary pre-orders, however, would be worth pursuing from a "constructive" point of view, when studying complex legal codes as structures that emerge from the order-theoretical composition of P-sequences viewed as basic building blocks of normative codes. In exploring this, our analysis of betterness dynamics should be pushed further by exploring further operations on P-sequences and orders suggested by moral and legal reasoning.

Finally, as we have seen in Chapters 5, 8, preference tends to occur entangled with informational attitudes like knowledge and belief. The same is true in deontic and legal settings, where information can be crucial to our rights and duties (cf. [149]).[12] Thus, just like in other parts of this book, deontic betterness dynamics must eventually interact with the dynamics of acquiring relevant information, and its resulting changes in attitudes such as knowledge and beliefs.

11.6 Appendix: A Semantic Excursion on Imperatives

11.6.1 Motivation

Imperatives occur ubiquitously in linguistic communication between individuals. To act successfully in society, we have to fully understand their meaning, as they lead to an implementation of actions. And as far as the dynamics of normative reasoning is considered, imperatives are the linguistic triggers that change our preferences, or other structures crucial to our rights and obligations. Thus, issues in the semantics of natural languages may be close to issues of normativity and agency in general.

Dynamic semantics and dynamic logic Logical studies of imperatives have been pursued for a long time, but starting from the 1990s, new frameworks have appeared. Here is one that is especially relevant to what we have done in this chapter.

Following the idea that "You know the meaning of a sentence if you know the change it brings about in the cognitive state of anyone who wants to incorporate the information conveyed by it", a dynamic update semantics was proposed for natural language default rules in [192]. This style of thinking was later applied to imperative expressions in [118, 139]. This approach falls within the influential tradition of "dynamic semantics" for natural language, where expressions of familiar languages get new dynamic meanings: cf. [150] for further details.

[12] Whether we are guilty of neglect, say, may depend on whether we took steps to make sure that we knew the relevant facts – even though no one is held morally to be omniscient.

By now, another style is also available. The *DEL* approach that adds explicit modalities for speech acts to a classical base logic was introduced in Chapter 2 for use in logics of agency. It, too, has been adopted to deal with the speech act of commands. Notably, [199] and [200] introduced a new dynamic action of "commanding" into static deontic logic, and dealt with imperatives in the framework of dynamic deontic logics. In this book, we have no view on which of these approach better fits natural language (if one has to choose at all). But we do want to point out that, construed either way, the perspective of this chapter has a contribution to make.

A case study: conflicting commands So far, the main purpose of both mentioned approaches has been to understand the meaning of one *single* imperative. Not much attention has been paid to the larger setting of imperative discourse, where there my typically be *conflicting commands*. Consider the following two examples, the first of which is a variation of Yamada's motivating example in [199]:

Example 11.13 Suppose you are reading an article on logic in the office you share with your two bosses, a_1 and a_2 and a few other colleagues. We assume that a_2 is of a higher rank. While you are reading, the temperature of the room rises, and it is now above 30 degrees Celsius. There is a window and an air conditioner. You can open the window, or turn on the air conditioner. You can also concentrate on the article and ignore the heat. Then, suddenly, you hear your boss a_1's voice. She commanded you to open the window. Right after that, a_2's voice comes up, saying he prefers that the window is not opened (This can be taken as an command "do not open the window"). What effects do those commands have on the current situation?

The next example shows the problem in an even more explicit manner:

Example 11.14 A general, a captain and a colonel utter the following sentences, respectively, to an agent b:

(1) The general a_1: Do A! Do B!
(2) The captain a_2: Do B! Do C!
(3) The colonel a_3: Don't do A! Don't do C!

It is clear that there are conflicting orders, w.r.t action A and C. Intuitively, instead of getting stuck, agent b will come up with the following plan after a deliberation: She should do A, do B, but not do C. b's reasoning rests on the following fact, the authorities of a_1, a_2 and a_3 are ranked as follows: $a_1 \gg a_3 \gg a_2$. According to this ranking, she can make her decision on which orders to obey, which orders to disobey. Thinking in terms of priorities and preference from Part IV, authorities amount to priorities, it will give us a preference over actions.

The above-mentioned frameworks cannot meaningfully handle such complex cases. The main problem is that their focus has always been on the *addressee*, not on the agents uttering the command. Inspired by the above model of priority-based preference, the aim of this section is to propose a solution to the problem of conflicting commands, in both mentioned frameworks.

What follows is just a small case study in dynamic semantics, relying heavily on the basic framework for imperatives put in place by the cited references.

11.6.2 Update Semantics for Conflicting Imperatives

A new update system is a tuple $(\mathscr{L}, \Sigma, [\cdot], A, \gg)$, where A is a finite set of agents, i.e. speakers. $\gg$ is a partial order on the set A. Intuitively, $\gg$ is an authority order of the speakers. Now we formulate the semantics in details, incorporating the authorities in the framework presented in the preceding.

Definition 11.15 (agent-oriented language $\mathscr{L}$) Let $\mathscr{Y}$ be the standard language of propositional logic. Define the set $\mathscr{L}$ of imperatives as follows:

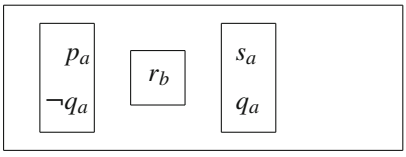
$$\mathscr{L} = \{!_a\varphi \mid \varphi \in \mathscr{Y}, a \in A\}$$

We see that imperatives are relative to specific speakers.

Definition 11.16 (force structures with authorities) Let L be the set of literals of $\mathscr{Y}$. Let $\mathsf{L}' = \{l_a \mid l \in \mathsf{L}, a \in A\}$. Let $\mathsf{B} = \{X \subseteq \mathsf{L}' \mid X \text{ is finite, and for any } l_a, l_b \in X, a = b\}$. Each $J \in \mathsf{B}$ is called a choice scope. Let $\mathsf{F} = \{X \subseteq \mathsf{B} \mid X \text{ is finite}\}$. Each $K \in \mathsf{F}$ is a force structure. The empty set $\emptyset$ is called the minimal force structure. Those force structures containing $\emptyset$ are called absurd ones.

Here is an example of force structures that involve authorities:

Example 11.17 Consider the structure $K_1 = \{\{p_a, \neg q_a\}, \{r_b\}, \{s_a, q_a\}\}$, which can be represented by the following picture:



We see that in force structures, literals are relative to specific speakers. Each literal can be viewed as an atomic imperative force. In this sense, we can say that imperative forces are relative to particular speakers.

Definition 11.18 (track with authorities) Let $K = \{X_1, \ldots, X_n\}$ be any force structure. We define tracks for K. For any X_i, let X_i' be the smallest set such that both p_a and $\neg p_a$ are in X_i' for any p_a occurring in X_i. $T = X_1'' \cup \cdots \cup X_n''$ is a track for K if and only if:

(1) $X_i'' \subseteq X_i'$ and $X_i'' \cap X_i \neq \emptyset$;
(2) For any p_a occurring in X_i, one and only one of p_a and $\neg p_a$ is in X_i''.

T is consistent if and only if there are no p_a and p_b such that both p_a and $\neg p_a$ are in T, but either $a \gg b \wedge b \gg a$ or $\neg(a \gg b) \wedge \neg(b \gg a)$.

T is resolvable if and only if there are no p_a and p_b such that both p_a and $\neg p_b$ are in T but either $a \gg b \wedge \neg(b \gg a)$ or $\neg(a \gg b) \wedge b \gg a$.

Note that a consistent track might not be resolvable, and also vice versa.

Definition 11.19 (imperatives and force structures) F is the set of force structures. Let K be any force structure. T^+ and T^- are two functions from $\mathsf{F} \times \mathscr{L}$ to F, which are defined in the following way:

$$
\begin{array}{ll}
(1) & T^+(K, !_a p) = \begin{cases} \{\{p_a\}\} & \text{if } K = \emptyset \\ \{X \cup \{p_a\} \mid X \in K\} & \text{otherwise} \end{cases} \\
 & T^-(K, !_a p) = \begin{cases} \{\{\neg p_a\}\} & \text{if } K = \emptyset \\ \{X \cup \{\neg p_a\} \mid X \in K\} & \text{otherwise} \end{cases} \\
(2) & T^+(K, !_a(\neg\varphi)) = T^-(K, !_a\varphi) \\
 & T^-(K, !_a(\neg\varphi)) = T^+(K, !_a\varphi) \\
(3) & T^+(K, !_a(\varphi \wedge \psi)) = T^+(K, !_a\varphi) \cup T^+(K, !_a\psi) \\
 & T^-(K, !_a(\varphi \wedge \psi)) = T^-(T^-(K, !_a\varphi), !_a\psi) \\
(4) & T^+(K, !_a(\varphi \vee \psi)) = T^+(T^+(K, !_a\varphi), !_a\psi) \\
 & T^-(K, !_a(\varphi \vee \psi)) = T^-(K, !_a\varphi) \cup T^-(K, !_a\psi)
\end{array}
$$

For any imperative $!_a\varphi$, $T^+(\emptyset, !_a\varphi)$ is its corresponding force structure.

Next, we define the notions of compatibility and harmony for force structures.

Definition 11.20 (compatibility) Let K_1 and K_2 be any force structures. We say that K_1 and K_2 are compatible if and only if

(1) For any track T_1 of K_1, there is a track T_2 such that $T_1 \cup T_2$ is consistent;
(2) For any track T_2 of K_2, there is a track T_1 such that $T_1 \cup T_2$ is consistent.

Definition 11.21 (harmony) Let K_1 and K_2 be any force structures. We say that K_1 and K_2 are harmonious if and only if

(1) For any track T_1 of K_1, there is a track T_2 such that $T_1 \cup T_2$ is resolvable;
(2) For any track T_2 of K_2, there is a track T_1 such that $T_1 \cup T_2$ is resolvable.

In what follows, we only consider the simplest case, namely, those imperatives whose force structures contain single choice scopes. As we know, force structures containing single choice scopes only has one track.

Consider two arbitrary force structures K and K'. Let $\Sigma = \{K_1, \ldots, K_n\}$, where each set K_i satisfies the following condition:

(1) $K_i \subseteq K \cup K'$;
(2) The track of K_i is resolvable;
(3) For any $K_i' \subseteq K \cup K'$, if $K_i \subset K_i'$, then the track of K_i' is not resolvable.

Definition 11.22 (update function) Define the function U as follows: $U(K, K') = K''$, with K'' the greatest element of Σ. The update function $[\cdot]$ is defined as follows:

$$
K[!_a\varphi] = \begin{cases} U(K, T^+(\emptyset, !_a\varphi)) & K \text{ and } T^+(\emptyset, !_a\varphi) \text{ are consistent and compatible} \\ \{\emptyset\} & \text{otherwise} \end{cases}
$$

With all our techniques ready, we come back to Example 11.14:

(1) The general a_1: Do A! Do B!
(2) The captain a_2: Do B! Do C!
(3) The colonel a_3: Don't do A! Don't do C!

Suppose that starting point of update is $\emptyset$. After the captain's orders, the force structure of the agent b is this: $\{\{A_{a_1}\}, \{B_{a_1}\}, \{B_{a_2}\}, \{C_{a_2}\}\}$. It can be verified that $\{\{A_{a_1}\}, \{B_{a_1}\}, \{B_{a_2}\}, \{C_{a_2}\}\}[!_{a_3}(\neg A)][!_{a_3}(\neg C)] = \{\{A_{a_1}\}, \{B_{a_1}\}, \{B_{a_2}\}, \{C_{a_2}\}\}$ $[!_{a_3}(\neg C)] = \{\{A_{a_1}\}, \{B_{a_1}\}, \{B_{a_2}\}, \{\neg C_{a_3}\}\}$. The only track of the force structure $\{\{A_{a_1}\}, \{B_{a_1}\}, \{B_{a_2}\}, \{\neg C_{a_3}\}\}$ is consistent and resolvable. This implies that agent b has to do A and B, and he is forbidden to do C. This is precisely what happens in real life. We take orders, with a consideration on the resources of information. This ends our investigation in update semantics.

However, using the same ideas, we can introduce the authorities into the *DEL*-style dynamic deontic logic. Essentially, the sequence of authorities gives rise to an ordering or a choice over the commands that the agent gets. Accordingly, the agent updates with right information in the right order. In what follows, we will adapt such an idea to Yamada's framework, providing some basic definitions. We will leave the systematic study to other occasions.

11.6.2.1 Dynamic Deontic Logics for Conflicting Commands

Definition 11.23 (static deontic language) Let $p \in \Phi$, a set of proposition letters, and $i, j \in N$, a finite set of agents, $\gg$ is a partial order over N. The static deontic language is given by:

$$\varphi := \bot \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid U\varphi \mid O_{(i,j)}\varphi$$

Intuitively, the formula of the form $O_{(i,j)}\varphi$ means that it is obligatory upon the agent i with respect to the authority j that i should see to it that φ. Note that the set N is now an ordered set of agents.

Definition 11.24 (semantics) A deontic model is a tuple $\mathfrak{M} = (S, \{\smile_{(i,j)} \mid i, j \in N\})$, with S a set of possible worlds, V a valuation for proposition letters. Moreover, $\smile_{(i,j)}$ is an arbitrary relation over the worlds.

Definition 11.25 (truth conditions) Given a deontic model $\mathfrak{M} = (S, \{\smile_{(i,j)} \mid i, j \in N\})$, and a world $s \in S$, we define the relation $\mathfrak{M}, s \models \varphi$ (formula φ is true in $\mathfrak{M}$ at s) by induction on φ:

$$\mathfrak{M}, s \models O_{(i,j)}\varphi \quad \text{iff} \quad \text{for all } t \in S \text{ such that } s \smile_{(i,j)} t, \mathfrak{M}, t \models \varphi.$$

In order to talk about changes that acts of commanding bring about, we extend the deontic language by adding action modalities to it:

Definition 11.26 Let $p \in \Phi$, a set of proposition letters, and $i, j \in N$, a finite set of agents, $\gg$ is a partial order over N. The dynamic deontic language is given by:

$$\varphi := \bot \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid U\varphi \mid O_{(i,j)}\varphi \mid [\pi]\varphi$$
$$\pi := !_{(i,j)}\varphi$$

A formula of the form $[!_{(i,j)}\varphi]\psi$ means that after an act of commanding addressed to an agent i by an authority j to the effect that i should see to it that φ, ψ holds.

Definition 11.27 (deliberation function) Consider a finite partially-ordered set of agent A, and any atomic command p and its negation $\neg p$ issued by different authorities, we define a deliberation function for the addressee i:

$$f(p_{(i,j)}, \neg\, p_{(i,k)}) = \begin{cases} p_{(i,j)} & \text{if } j \gg k; \\ \neg p_{(i,k)} & \text{if } k \gg j. \end{cases}$$

This function describes the process of agent's resolving conflicts and reaching a harmony ordering over commands she has received. Once this process is done, the rest of her work is routine. In the following, we give the truth definition for the relevant dynamic modality:

Definition 11.28 Given a deontic model $\mathfrak{M} = (S, \{\smile_{(i,j)} \mid i, j \in N\})$, and a world $s \in S$, the truth definition for formulas is as before, but with one new clause for the action modalities:

$$\mathfrak{M}, s \models [!_{(i,j)}\varphi]\psi \quad \text{iff} \quad \mathfrak{M}_{!_{(i,j)}\varphi}, s \models \psi,$$

where $\mathfrak{M}_{!_{(i,j)}\varphi}$ is a deontic model obtained from $\mathfrak{M}$ by replacing $\smile_{(i,j)}$ with $\smile_{(i,j)} - \{(s,t) \mid \mathfrak{M}, t \models \neg\varphi\}$.

11.6.2.2 Conclusion

We have shown how ideas from the richer representation style of this chapter can be used to enrich existing dynamic semantics for natural language. In particular, the preference-changing events studied in this book often come couched in linguistic expressions whose meaning is clearly relevant to our analysis of agency. Even so, our aim in presenting a small case study of imperatives has not been to make grand claims about the proper treatment of natural language. We only established this: Ideas can flow across from our dynamic preference logics to what initially may look like quite different frameworks.

Chapter 12
Games and Actions

12.1 Introduction

Our main focus in this book has been the representation of preferences of one agent, their entanglement with informational attitudes like beliefs, and the various events that dynamically change these informational and evaluative attitudes. But like all aspects of agency, such single steps are the building blocks for something larger, viz. interaction between different agents over time. For instance, a learning process is typically a long-term history where an agent interacts with a source of information, and where both knowledge, beliefs, and goals may change over time. At its most pregnant, this longer-term interactive aspects occurs in *games*, where many of our main concerns occur very concretely – including significant meetings between information and evaluation.

Consider this simple example from the literature (cf. [28]):

Example 12.1 (a simple game) There are two players, Abelard (A) and Eloise (E), with possible actions are "going left" and "going right". Their payoff is represented by the pairs at the leaves of the game tree. E.g. $(1, 0)$ expresses that A gets 1 and E gets 0. The game as drawn starts with A's move, then follows E's move, unless A goes left, ending the game at once (Fig. 12.1).

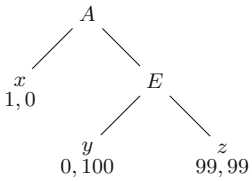


Fig. 12.1 A simple game

What would A do first? The famous *Backward Induction* procedure[1] claims that A would go left, after which the game is over. Its reasoning is as follows. After comparing (0, 100) and (99, 99), A *believes* that E would go left when it becomes

[1] For more information about Backward Induction and its history, see [147].

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_12, © Springer Science+Business Media B.V. 2011

her turn, since that choice will give her the best pay-off. In that case, *A* gets 0. So *A* will go left at the beginning, to make sure he gets 1 which is *better* than 0.

This game scenario vividly shows us how players form beliefs about the behavior of others, and then choose their actions based on their further beliefs and preferences in games. What we see then is a concrete scenario for many of notions so far, preference and belief entangled, but now mixed with a third ingredient of *action*. In this chapter, we will investigate how the preference logics of this book interface with game theory. We will find many promising links, though developing a full-fledged merged theory would be beyond the scope of this book. There is a large and fast-growing body of research on logic and games, for which references will be given in what follows. The following account of some relevant themes draws heavily on [46],[2] but eventually, I turn to my own considerations in the light of this book.

This chapter proceeds in the following stages. First, in Section 12.2, I briefly explain how modal preference logic occurs in strategic games, with strategy profiles now serving as concrete structured worlds, instead of the abstract worlds used in Chapter 3. This illustrates at once several themes from this book, especially, the entanglement of preference and epistemic notions, and the need for extended languages, in defining such "entangled" notions as best response and Nash Equilibrium. In Section 12.3, I look at games in extensive form and illustrate how preference mixes with action to yield a logical definition for the Backward Induction solution. The key notion in this analysis is that of *rationality*, an entangled notion of best action given one's beliefs about the future course of the game. Next, in Section 12.4, I review some recent proposals for a dynamic analysis of game solution. The key feature here are repeated announcements of hard information of rationality until a stable limit is reached, or alternatively, repeated soft upgrades making rationality more plausible.

Next, in Section 12.5, I connect up more closely with the main concerns of this book. First I consider some new directions that my preference logics might learn from the focus on games. My examples will be game-theoretic scenarios behind the examples in earlier chapters, new forms of set lifting for preference, analogies between action and belief, logics with preference directly on actions rather than worlds, and diversity of agents. After that, in Section 12.6, I reverse the direction, and discuss how the main concerns in this book might affect the logical study of games. In particular, I make some proposals concerning the role of *preference change*, and *priority structure* in games. Generalizing from the case of games, I conclude with some general thoughts on long-term temporal perspective in preference modeling.

12.2 Preference Logic in Strategic Games

We start with a simple point of departure for much recent literature (cf. the survey [106]). Games are natural models for many of the logical languages found in this book: in particular, epistemic, doxastic and modal preference logics.

[2] I thank the authors for their permission to use relevant material from their survey.

We first briefly consider strategic games as a simple setting where our earlier abstract possible worlds models acquire concrete structure.

Definition 12.2 (strategic game) A strategic game G for a set of n players N consists of the following two components:

(1) a nonempty set A_i of actions for each $i \in N$,
(2) a utility function or preference ordering on the set of outcomes.

For simplicity, one often identifies the outcomes of G with the set $S = \Pi_{i \in N} A_i$ of *strategy profiles*. Given a strategy profile $\sigma \in S$ with $\sigma = (a_1, \ldots, a_n)$, σ_i denotes the i-th projection (i.e., $\sigma_i = a_i$) and σ_{-i} the choices of all agents except agent i: $\sigma_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$.

Game models and modal logics Now, from a logical perspective, it is natural to treat the set S of strategy profiles as "possible worlds" carrying three relations[3]:

- $\sigma \geq_i \sigma'$ iff player i weakly prefers the outcome σ over the outcome σ',
- $\sigma \sim_i \sigma'$ iff $\sigma_i = \sigma'_i$: This epistemic relation represents player i's "view of the game" when a player has chosen her move, but does not know that of the others (cf. [31]),
- $\sigma \approx_i \sigma'$ iff $\sigma_{-i} = \sigma'_{-i}$: this relation of "action freedom" gives the alternative choices for player i when the other players' actions are fixed (cf. [168]).

Such "modal game models" $\mathfrak{M} = (S, \{\sim_i\}_{i \in N}, \{\approx_i\}_{i \in N}, \{\geq_i\}_{i \in N})$ can obviously interpret modal languages of the kind used in this book for knowledge and preference, while even adding one for action. Here is how, in an obvious notation:

- $\sigma \models \Box\varphi$ if for all σ' if $\sigma \sim_i \sigma'$ then $\sigma' \models \varphi$.
- $\sigma \models \boxminus\varphi$ if for all σ' if $\sigma \approx_i \sigma'$ then $\sigma' \models \varphi$.
- $\sigma \models \langle \geq_i \rangle \varphi$ if there exists σ' such that $\sigma' \geq_i \sigma$ and $\sigma' \models \varphi$.

Logical issues: definability and complexity The logic of these game models immediately illustrates some of our earlier key issues, but now in a concrete setting. For instance, the abstract "entanglement" of information and evaluation in Chapter 5 now shows in the connections between the three relations in the above models, as one moves across rows and columns of a game matrix. Here are two concrete issues which illustrate how our earlier concerns tie in with logical properties of the system.

Preference and best response To this setting of available actions, information, and freedom, the preference structure of our book adds further interesting features. One benchmark for modal game logics has been the definition of the strategy profiles that are in *Nash Equilibrium*, or rather, the usual notion of *best response* for a player. The latter is not definable in our modal language so far. The bisimulation-based argument is similar to one we gave for epistemically entangled preference in Chapter 5, and indeed, as we did there, we need to add an *intersection modality*

$$\mathfrak{M}, \sigma \models [\approx_i \cap >_i]\varphi \quad \text{iff} \quad \text{for each } \sigma' \text{ if } \sigma (\approx_i \cap >_i) \sigma' \text{ then } \mathfrak{M}, \sigma' \models \varphi.$$

[3] More abstract worlds might carry strategy profiles without being identical to them.

Then we can define best response for player i by means of the modal formula

$$\neg\langle \approx_i \cap >_i \rangle \top.$$

Thus, equilibrium notions in games exemplify the "deep entanglement" of preference with other notions that we identified earlier.

Logical validities What is the modal calculus of reasoning for strategic games? First, given the nature of our three relations, the separate logics are as earlier in our book: Modal **S4** for preference, and modal **S5** for epistemic outlook and action freedom, since the latter are clearly equivalence relations. What is of greater interest, and logical delicacy, is the *interaction* of the three modalities. For instance, the following combination of two modalities makes φ true in each world of a game model:

$$\Box \boxminus \varphi$$

Thus, the modal game language has the earlier "universal modality" of our preference logics for free. Moreover, this modality can be defined in two ways, since

The equivalence $\Box \boxminus \varphi \leftrightarrow \boxminus \Box \varphi$ is valid in strategic game models.

This validity depends on the geometrical "grid property" of game matrices that

if $x \sim_i y \approx_i z$, then there exists a world u with $x \approx_i u \sim_i z$.

Pitfalls of complexity But now games show us something that has not really surfaced in our earlier discussions of entanglement. The above geometric grid condition looks like a pleasant pictorial feature of matrices, but its logical effects are delicate. The general logic of such a bi-modal language on grid models is not decidable, and not even axiomatizable.[4] Thus, from our point of view, strategic games have a bitter-sweet flavour. They provide concrete instances of the abstract models in this book, but their regular structure may also lead to very complex combined logics of agency.

Our main point with this warm-up discussion is just this. Simple strategic games are a concrete model for the logics in this book, with vivid intuitions behind them. They also show how combining preference with natural other features of agency may result in quite rich and unpredictable logical systems.[5]

12.3 Preference Logic in Extensive Games

Preference plays even more substantially in the setting of *extensive games* that record the actual course of play. We demonstrate this with a recent case study of *Backward Induction*, a famous procedure for solving extensive games that was

[4] cf. [144] and [45] for formal details.

[5] There is a much more literature on these topics. cf. [25, 26, 56, 106], and [137].

already introduced briefly at the beginning of this chapter. We start with static logics, and subsequently, bring in the information dynamics of Chapters 2 and 4.

12.3.1 Dynamic Logic of Actions and Strategies

An extensive game mixes action structure and preference structure. One obvious language for describing the former is *propositional dynamic logic* (*PDL*), a system used already in Chapters 4 and 10. We recall some basics.

Let Act be a set of primitive actions. An *action model* is a tuple $\mathfrak{M} = (S, \{R_a \mid a \in \mathsf{Act}\}, V)$ where S is an abstract set of states, or stages in an extensive game, and for each $a \in \mathsf{Act}$, $R_a \subseteq S \times S$ is a binary *transition relation* describing possible transitions from states s to s' by executing the action a.

On top of this atomic repertoire, the tree-like structure of extensive games also supports complex action expressions, constructed by the standard regular operations of "indeterministic choice" ($\cup$), "sequential composition" ($;$) and "unbounded finitary iteration" (*, also called Kleene star):

$$\alpha \ := \ a \mid \alpha \cup \beta \mid \alpha; \beta \mid \alpha^*.$$

The key dynamic modality $[\alpha]\varphi$ then says that "after the move described by the program expression α is taken, φ is true":

$$\mathfrak{M}, s \models [\alpha]\varphi \ \text{ iff } \ \text{for each } t, \text{ if } s R_\alpha t \text{ then } \mathfrak{M}, t \models \varphi.$$

PDL has been used to define solution concepts on extensive games by many authors (cf. [105, 106] and [25], and for defining explicit strategies in [30]).

12.3.2 Adding Preferences: The Case of Backward Induction

As before, the complete picture must bring in players' preferences on top of *PDL*, again along the lines of our modal preference logic of Chapter 3.

To show how this works, we consider a standard pilot example: the Backward Induction (*BI*) algorithm for finite games, as in our initial Example 12.1. In its usual format, this procedure marks nodes of an extensive game tree with values for the players encoding the best that they can guarantee by appropriate play henceforth, and assuming that all others do likewise. For better fit with our logical setting, we will work with a relational version of the *BI* algorithm. In ordinary parlance, a strategy restrict one's choices in some way, not necessarily unique – like a plan. Technically, this makes a strategy a sub-relation of the total *move* relation in a game.

Relational BI First, mark all moves as *active*. Call a move *a dominated* if it has a sibling move all of whose reachable endpoints via active nodes are preferred by the current player to all reachable endpoints via *a* itself. We work in stages: *At each stage, mark dominated moves in the* $\forall\forall$ *sense of preference as* passive*, leaving all others active.* Here, the "reachable

endpoints" by an active move are all those that can still be reached via a sequence of moves that are still active at this stage.

The driving idea behind relational *BI* is the following form of *Rationality*:

I do not play a move when I have another move whose outcomes I prefer.

This procedure is cautious, in that players avoid "strictly dominated moves".[6]

Dominance and lifted set preference There is an interesting connection here with our earlier notion of generic preference between sets of worlds (Chapter 3). Preferences between moves that stand for *different sets of outcomes* call for a notion of *lifting* the given preference on end-points of the game to sets X, Y of end-points. And the way we do this matters. The quantification pattern used for lifting in the above algorithm was the $\forall\forall$ clause (cf. [39]) that

$$\forall x \in X \forall y \in Y \; x <_i y.$$

But other stipulations make sense. For instance, our running example in Chapter 3 was rather the following quantifier combination:

$$\forall x \in X \exists y \in Y \; x <_i y.$$

In game-theoretic terms, this says that we should choose a move with the highest maximal value that can be achieved. This is not the standard approach, however, which would rather try to choose a move guaranteeing the highest *minimal values*, expressed by the quantifier combination

$$\forall x \in X \exists y \in Y \; x >_i y.$$

But now, our earlier notions of set lifting acquire a very concrete meaning. They are not abstract options to be decided by philosophizing: They rather express *different legitimate styles of play*, either more cautious, or more risk-taking. One game might be played by agents following different policies of this sort. In other words, in principle, the technical diversity of Chapter 3 was not a drawback, but an advantage in realistic modeling of games.

12.3.3 Backward Induction in Preference-Action Logic

The power of combining dynamic logic with preference logic shows in this result from [44] defining Backward Induction:

Theorem 12.3 *For each extensive game form, the strategy profile σ is a backward induction solution iff σ is played at the root of a tree satisfying the following modal axiom for all propositions p and players i*[7]*:*

---

[6] Rationality is a sweeping form of "entanglement", a bridge law between what we know or believe about the outcomes of our actions and how we evaluate these.

[7] Here $move_i = \bigcup_{a \text{ is an } i\text{-move}} a$, turn_i is a propositional variable saying that it is i's turn to move, and *end* is a propositional variable true at only end nodes.

$$(turn_i \wedge \langle \sigma^* \rangle (end \wedge p)) \rightarrow [move_i] \langle \sigma^* \rangle (end \wedge \langle \geq_i \rangle p)$$

The meaning of the crucial axiom follows by a modal frame correspondence:

Fact 12.4 *A game frame makes* $(turn_i \wedge [\sigma^*](end \rightarrow p)) \rightarrow [move_i]\langle \sigma^* \rangle (end \wedge \langle \geq_i \rangle p)$ *true for all* i *at all nodes iff the frame has this property for all* i*:*

RAT: No alternative move for the current player i *guarantees outcomes via further play using* σ *that are all strictly better for* i *than all outcomes resulting from starting at the current move and then playing* σ *all the way down the tree.*

A picture from [82] reflects this notion of rationality, that we discussed before. This is how the entanglement of preference and action shows in a concrete setting (Fig. 12.2):

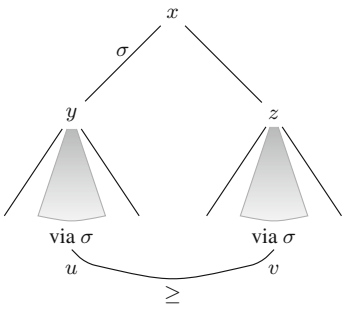


Fig. 12.2 Entanglement of preference and action

More formally, RAT is this *confluence property* for action and preference:

$$\text{CF} \qquad \bigwedge_i \forall x \forall y ((turn_i(x) \wedge x\ \sigma\ y) \rightarrow$$

$$(x\ move\ y \wedge \forall z (x\ move\ z \rightarrow \exists u \exists v (end(u) \wedge end(v) \wedge y\ \sigma^*\ v \wedge z\ \sigma^*\ u \wedge u \leq_i v)))$$

Now, a simple inductive proof on finite game trees shows for our algorithm:

Theorem 12.5 *BI is the largest subrelation* S *of the move relation in a game satisfying (i)* S *has a successor at each intermediate node and (ii)* S *satisfies CF.*[8]

This concludes our survey of recent static analysis of the logic behind extensive games. But in the spirit of this book, these leave something to be desired. A crucial feature remains implicit: The *dynamics* of deliberation and information flow that determine players' expectations beforehand, and their actual play as a game unfolds. We now turn to this.[9]

[8] The definition of S can be stated in a well-known logic of computation, viz. *first-order fixed-point logic LFP(FO)* [70]. The dissertation [82] has details on this way of defining game solution methods.

[9] The following section mainly follows the topics and results of [38, 82].

12.4 Solution Dynamics in Extensive Games

In a dynamic perspective, the main interest of a solution concept is the way in which its outcomes are reached, its "rational procedure". In particular, Backward Induction is really also a procedure for creating players' *expectations*, and these involve an entangled mixture of beliefs and preferences, as we have studied in Chapters 5 and 8. This will now be given a more precise sense.

Knowledge and iterated public announcement of rationality An early dynamic take on Backward Induction was proposed in [31]. The process of pre-play deliberation is modeled there as repeated announcements of the assertion that

> *rat*: players are "rational" in the sense of *never playing a strictly dominated move* in our earlier ∀∀-sense of set preference.

This proposition *rat* may be true or false at nodes of a game tree, and hence, every time it is publicly announced, the extensive game may get smaller, in the manner of our public announcement logic *PAL* of Chapter 2. Iterated announcements then produce a shrinking nested sequence of models, and this sequence must reach a *limit* $(!rat, \mathfrak{M})^{\#}$, a first model where announcing the rationality proposition *rat* no longer rules out any nodes in the game tree.

Theorem 12.6 *In any game* $\mathfrak{M}$, $(!rat, \mathfrak{M})^{\#}$ *is the actual subtree computed by BI.*

Thus, the actual *BI* play is the limit sub-model, where *rat* holds throughout. In a term from the literature, Rationality is a "self-fulfilling" proposition: Its announcement eventually makes it true everywhere. In this way, games add an interesting social twist to our analysis of preference and action for individual agents.[10]

Belief and iterated plausibility upgrade However, as we said, Backward Induction creates expectations, i.e., beliefs about the future course of a game. Thus, the *BI* procedure does not eliminate nodes of the initial game, but endows them with progressive expectations on how the game will proceed. This is the plausibility dynamics that we studied in Chapter 3, now performing a *soft announcement* of the proposition *rat*, where the appropriate action is our earlier radical upgrade on end nodes of the game, viewed as completed histories or worlds:

Example 12.7 (the BI procedure with soft upgrades) We start with all endpoints of the game tree incomparable. Next, at each stage, we compare sibling nodes, using the following notion of belief-entangled preference:

> A move x for player i dominates a sibling move y *in beliefs* if the *most plausible* end nodes reachable after x along any path in the whole game tree are all better for the active player than all the *most plausible* end nodes that are reachable in the game after y.

Rationality* (rat^*) then says that no player plays a move that is dominated in beliefs. Now we perform what is essentially a radical upgrade $\Uparrow rat^*$:

---

[10] There are also "self-refuting" propositions, becoming false everywhere in the limit model of their repeated announcement. This happens, for instance, with the repeated ignorance assertions in the Muddy Children puzzle (cf. [73]).

If a game move x dominates move y in beliefs, make all end nodes reachable from x more plausible than those reachable from y, while keeping the old order inside these zones.

This changes the plausibility order, and hence the pattern of dominance-in-belief, so that iteration makes sense. Consider our initial Example 12.1 once more. Here are the update stages, with letters x, y, z standing for the end nodes of the game:

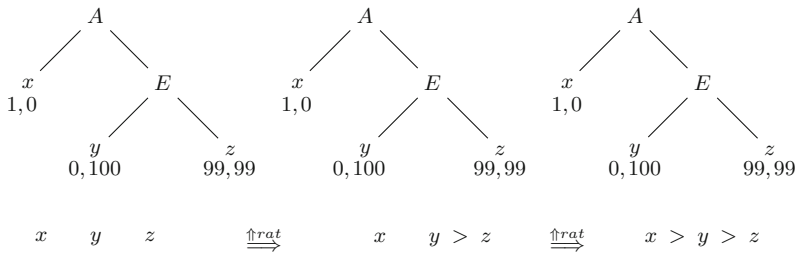


Fig. 12.3 Changes in plausibility relations

In the first tree, going right is not yet dominated in beliefs for A by going left. rat^* only has bite at E's turn, and the first upgrade makes (0, 100) more plausible than (99, 99). After this upgrade, going right has become dominated in beliefs, and a new upgrade takes place, making A's going left most plausible. Here is a general result proved in [32, 38]:

Theorem 12.8 *The Backward Induction strategy is encoded in the final plausibility order for end nodes after iterated radical upgrade with rationality-in-belief.*

Thus, the algorithmic view of Backward Induction and its procedural doxastic analysis in terms of forming preference-entangled beliefs amount to the same thing.

While the literature that we have followed here is after a dynamic analysis of game solution, the main points for us are about the role of preference. One is how, again, the abstract preference logics of our earlier chapters now acquire a concrete meaning in simple action scenarios. The other point is that preference becomes a much more exciting driving force in logical systems when it is *combined* with other logical notions. But as we will see in a moment, there may be a price to pay.

Logic of preference and action: the complexity of entanglement The game logics of preference and action that result from the above analysis are not our main concern. We merely note that, as with strategic games, interesting questions arise about complexity of the combined systems (cf. [82]).

First consider an analogy with combined logics of action and knowledge. In this area, many authors have noted that apparently harmless "bridge assumptions" of Perfect Recall for agents with flawless memory make the validities undecidable, non-axiomatizable, and sometimes even of much higher complexity. The reason is that these assumptions generate commuting diagrams for actions *move* and epistemic uncertainty $\sim$ satisfying the following "confluence property":

$$\forall x \forall y((x \; move \; y \wedge y \sim z) \rightarrow \exists u(x \sim u \wedge u \; move \; z)).$$

These patterns then serve as the basic grid cells in encodings of complex "tiling problems" in the logic. Thus, the logical theory of games for players with perfect memory is more complex than that of forgetful agents (cf. [45]).

But grid-like patterns may also arise for other reasons. Recall the above non-epistemic property *CF* of *rationality*, that mixed action and preference in the style of Fig. 12.3. This, too, is a sort of geometrical confluence. Can it be that Rationality, an entanglement property for preference and action meant to make behaviour simple and predictable, actually makes its logical theory complex?

12.5 From Games to Preference Logic

We have presented some recent work on logic of games as a concrete illustration of how two major topics of this book fare in a concrete setting: *preference structure* and *update dynamics*. In doing so, we encountered quite a few of our earlier themes. We start with some obvious connections.

Game versions of preference scenarios For a start, many of our earlier preference scenarios could naturally be cast as multi-agent games. For instance, deontic reasoning involves preferences of interacting moral actors and moral authorities, or more practically, buying a house involves an interplay of buyers and sellers with not necessarily the same preferences, and so on.[11] Games always seem on the horizon as the next level of realism in modeling a scenario.

Varieties of set-lifted preference We encountered our technical theme of *set lifting* for betterness on worlds to generic preferences between propositions as a concrete issue of modeling different types of player, all rational, but differing in how they construe preference between sets of outcomes. The variety of options in Chapter 3 now became a legitimate variety for different players of a game.

Games even suggest new notions of set lifting beyond what we considered earlier. One example would be to just conjoin earlier stipulations, as in preferring moves with both the highest maximum outcome and the highest minimum outcome. But players can go even further, looking more deeply into the structure of the sets being compared. In Chapter 5, we defined doxastic preference for set Y over set X as a comparison between only the *most plausible* members of Y and X. But now, let us go on, and look beyond this 'minimal fringe:

Example 12.9 (deep-level comparison) Let A start the following game with two moves. Going left gives value 0, going right yields a turn for E, where she has two moves, one of which makes her better off, while A gets 0, and one of which makes E worse off, but A gets 1 (Fig. 12.4):

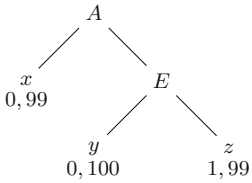


Fig. 12.4 Deep-level comparison

[11] Game-theoretic solution procedures have been applied to moral deliberation in [136].

Intuitively, A would choose going right toward E, even though he do not believe that it will improve his outcome. At least, it has a possibility of getting more, and he loses nothing compared to his move ending just with 0.

Here the comparison is between, first, the most plausible objects in the sets, and if that leads to indifference, then between the next most plausible objects in the sets, and so on. Finding the logic of this "deep comparison" between sets seems an interesting open problem in the spirit of Chapter 5.[12]

Action, preference, and revealed belief Game solution methods like *BI* also illustrate our earlier concern with links between action, preference and belief. As suggested in [20], a notion of "best action" is at the same time one of *revealed belief and preference*. Technically, this is because relational strategies correspond with plausibility relations. In particular, each sub-relation R of the total *move* relation in an extensive game induces a total plausibility order $ord(R)$ on endpoints of the game:

> $x\ ord(R)\ y$ iff, looking upwards at the first node z in the game tree where the histories of the endpoints x, y diverged, if x was reached via an R move from z, then so is y.

In terms of our earlier upgrade analysis, the precise facts are these:

Fact 12.10 *There is a one-to-one correspondence between sub-relations of the total* move *relation of a game and connected orders of the leaves.*[13]

Fact 12.11 *In terms of inductive computation stages, for each game tree* $\mathfrak{M}$ *and any* k, $rel((\Uparrow rat^*)^k, \mathfrak{M}) = BI^k$.

Thus, creating belief structure via upgrade is tightly correlated with determining what are our *best actions*.

We conclude with two more radical ways in which games may affect preference logic as presented in this book.

Preference directly on actions? Let us reconsider the very *locus* of preference in betterness relations. Should these relations run between worlds, as we have assumed, and then between sets of worlds as possible outcomes of actions, or could they also run directly *between actions*? The choice is reminiscent of one in deontics, between "deontology", where actions themselves can be good or bad, versus "utilitarianism", where actions are only good or bad in terms of how we evaluate their outcomes. Backward Induction, though starting from betterness on endpoints, derived how moves can be better than other moves. But in practice, one often just wants to know these best moves, and thus, it would be interesting to also put betterness directly between actions. In fact, [141] and [172] are precedents for this in the computational literature on agency.[14]

---

[12] Deep comparisons are reminiscent of how one computes a probabilistic *expected value*.

[13] This has a technical proviso: These leaf orders must be "node-compatible": cf. [82].

[14] Concretely, one might do this as follows, taking a cue from the *DEL* event models in Chapter 2. Start with models for *PDL* with preference between states. This interprets a standard dynamic

Agent diversity Here is a second new perspective that has been only an undercurrent in this book. One clear trend in logical analysis of games might be called *agent diversity*. In line with [46, 179] argues that game solution crucially needs assumptions about belief revision policies and other features of players.[15] There is no need to assume that players' styles of behavior are all the same. This theme of diversity fits well with the options that we have seen for set-lifted preferences: as naturally different styles of behavior, more cautious or more greedy. And it fits even more with the dynamic aspects of this book, where plausibility or betterness upgrades can come in many kinds, allowing agents great differences in behavior. Agent diversity has been studied more systematically in [41] and [132], looking at agents' powers of observation, inference, belief revision, and introspection.[16] Doing full justice to this important phenomenon is beyond the scope of this book.

12.6 From Preference Logic to Games

We have now seen two things in this chapter: Games form a natural model for the notions and concerns of this book, and also, they suggest new issues about the latter. Now we reverse the perspective, and ask whether the main themes of this book might have something of their own to add to current logical studies of games. We will discuss two main examples, having to do with our main themes of *preference change* and *priority structure*.

12.6.1 Preference Change in Games

We start with the simplest scenario of information dynamics in Chapter 2.

12.6.1.1 Game Change by Informational Events

In the preceding sections, we have already seen how the information dynamics of our book makes sense for games. Still, these examples, like the analysis of the Backward Induction procedure, may be considered a sort of "off-line" deliberative

---

language plus a preference modality. Now let actions by themselves – or better, *single transitions* between states as in Arrow Logic (cf. [23]) – form a modal model, with a binary betterness relation $\leq$ between transitions that can have unary atomic properties $p, q, \ldots$ The matching double modal language now also has formulas expressing properties of actions. One useful modal operator $END\varphi$ would say about a transition (arrow) a that the state-formula φ holds at the end-point of the a. And a second preference modality $\langle < \rangle \psi$, now on actions, will say that transition a sees a better transition b satisfying ψ. The logic will now also have axioms relating state and action modalities, depending on how one sees between preferring actions and preferring their outcomes. Backward Induction was one way of creating such links, but there may be others.

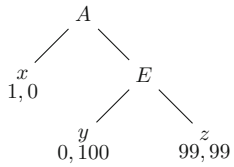
[15] Their slogan is that we need a "Theory of Play" going beyond game theory.

[16] Quantitative forms of diversity occurred in Chapter 6 with weighing rules for past experience and current observation. It is worthy of mentioning that [3] and [4] studied behaviors of resource-bounded agents.

dynamics about games, not internal "on-line dynamics" for players as the game proceeds (cf. [20] for this distinction). Still the systems in this book can also serve the latter.

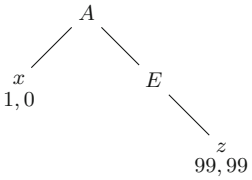
An early real game-changing scenario occurs in [148] on *announcing intentions*, where an agent says in the course of a game that she will restrict her behavior to only certain moves or strategies. This public announcement transforms the current game tree into a smaller one. Another example is the analysis of *promises* in [28], where again an important game transformation takes place under public announcement.

Example 12.12 (making a promise) In our running game example



Now E can promise A that "I will not go left when you have gone right." In the style of Chapter 2, this public announcement changes the game into (Fig. 12.5)

Fig. 12.5 After making a promise



Now, we can both be assured of getting an outcome of 99, as opposed to the meagre outcome 1, 0 predicted by standard *BI*.[17]

Game change in dynamic logic The two cited papers show how standard dynamic-epistemic logics apply in this setting, given a suitable static language over game models which describes players' moves, preferences, and beliefs:

Theorem 12.13 *There is a complete logic of public announcements over extensive games of perfect information which consist of a standard static base logic plus a complete set of reduction axioms for the announcement modalities over the relevant move and preference modalities of the game language.*

As an illustration, here is the reduction axiom for making a move a:

$$\langle !\varphi\rangle\langle a\rangle\psi \leftrightarrow (\varphi \wedge \langle a\rangle\langle !\varphi\rangle\psi).$$

Reduction axioms for preference, knowledge and belief are as in Chapters 2, 4.[18]

[17] Interestingly, from a deontic point of view, we made both players better off by restricting the freedom of one. Conversely then, an increase in freedom is not always a good thing.

[18] An interesting further issue is how public announcement changes effects of complex strategies in a game [30]. Consider the *strategy modality* $\{\sigma\}\psi$ saying that playing σ always leads to end points

12.6.1.2 Game Change by Evaluational Events

Can games also change by the acts of preference change that we have studied in this book? This seems less obvious, since one might think that a game only arises when all this has stabilized. For instance, suppose that a player has two moves *a, b*, preferring the outcome of *a* to that of *b*. Now at the start, she performs a preference change to *b* over *a*. The obvious thing to do seems to just say that the real game is now another one, with the preferences switched (Fig. 12.6):

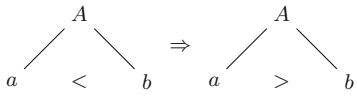


Fig. 12.6 Preference change inside a game

With preference changes inside a game, the transformation works on subgames. But one may also be interested in what actually happened during a play of the initial game, and then mind changes as to preference should be recorded explicitly. These changes could be of many forms.

Changing one's preferences post facto Suppose that an agent preferred the result of move *a* over that of move *b*, but she has in fact chosen *b*. She may now change her preference, and say "it was the best after all". This could be considered a form of internal *rationalization*. Conversely, the agent might choose the preferred move *a*, but afterwards, reverse her preference: "the grass is greener on the other side" once it has become unattainable.[19]

But of course, the preference change need not always be first-person. Rationalization is also a process undertaken by others, to make sense of observed behavior. Here is a folklore result, that we discuss in a version from [28]:

Rationalizing actions by preferences Rationality in the sense of decision theory or Backward Induction is very tenacious. One reason is its role, not in predicting human behavior, but in rationally *reconstructing* it. Suppose that your preferences between the outcomes of some game are not known. Then one can always ascribe preferences to you that make your actions rational. Here is a way of doing this.

Let a finite two-player extensive game G specify A's preferences, but not E's. The strategies σ_A, σ_E for playing G are given, yielding an expanded game model $\mathfrak{M}$. Now, when can A rationalize the observed behaviour σ_E to make the two strategies the Backward Induction solution of the game? What is clearly necessary here is a certain minimal quality of A's own moves, in fact, a form of our earlier Rationality assumption in *BI*:

> A's strategy chooses a move that is not strictly dominated in outcomes by another available move of his, given that the players follow σ_A, σ_E.

---

satisfying ψ. Here is a valid reduction axiom for reasoning about effects of strategies in the changed game: $\langle !\varphi\rangle\{\sigma\}\psi \leftrightarrow (\varphi \wedge \{\sigma|\varphi\}\langle !\varphi\rangle\psi$. Its notation $\{\sigma|\varphi\}$ refers to an obvious "relativization" of a *PDL* program σ to the submodel defined by φ.

[19] Impossibility can also be epistemic. I thought that I could reach the fruits and steal them. Now I cannot, and I think they have probably been sour anyway. This phenomenon has been discussed in decision theory and philosophy, cf. [72] on "sour grapes".

Let us call such a game "best-responsive" for A. The following is folklore:

Fact 12.14 *In any game that is best-responsive for A, there exists a preference relation for E among outcomes making the unique path that plays the two given strategies σ_A, σ_E against each other the Backward Induction solution.*

Proof The proof in [28] is a bit sketchy. Here is a more precise argument. A *bottom-up* procedure works as follows. A starts with final choices for players near the end of the game tree, assigning values reflecting preferences for E that make the given strategies the *BI* solution. The inductive step then works as follows. Assume that the subgames following all moves at the current top node have already been rationalized to yield the Backward Induction solution. There are two cases. (i) If the current node is A's turn, then best-responsiveness tells us that A's actual move is automatically optimal. (ii) If the current node is E's turn, then consider E's local σ_E move. We cannot be sure that its subgame guarantees a maximal value for E compared with the subgames for her other available moves. But we can change things to make this the case without changing anything essential, by *raising all values in the chosen subgame* by a suitably large number. This transformation is harmless:

> Raising all E's values of outcomes in a sub-game by a fixed amount N does not change the *BI*-solution, though it raises the total value by N.

□

Example 12.15 (rationalizing bottom-up) Here is an illustration of this procedure. Bold face lines are the given moves of the players, and numbers at the leaves indicate successive outcome values that are postulated *for* E. In the final step, an increase has been made as described above (Fig. 12.7).

Of course, utility numbers for E can be assigned in many ways to justify *BI*: Rationalizations are usually not unique. One can also reformulate the algorithm with qualitative betterness changes. For some purposes, this would even be simpler.[20]

Clearly this is just one rationalization scenario, but it shows the general spirit.

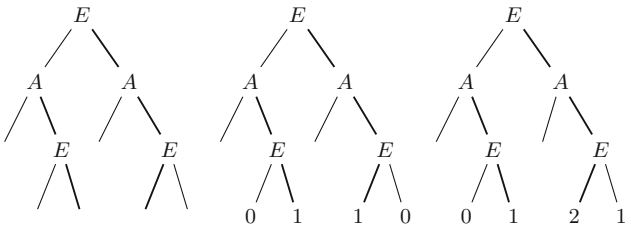


Fig. 12.7 Rationalizing bottom-up

[20] Indeed, given that Backward Induction is played on a certain set of preferences, one can *rationalize* the given preferences, in the sense of our earlier discussion, to new very special preferences that simplify the reasoning. For instance, in our initial running example, one might make the actual outcome the very best for all without changing the *BI* strategy.

Interestingly, one can also run the given procedure *top-down*, by making all outcomes of moves chosen by E better for her than the alternatives, and inside the sub-games resulting at each step, repeating this to creates finer preferences. The driving idea is this instruction:

> Make outcomes of actually chosen subtrees better for E than of those that were bypassed. This "upgrade" recipe is then iterated inside all subgame, until we reach the leaves.[21]

Actually, this procedure works better with successive relation changes for preference than with numbers, that need to be revised all the time to keep subgames in the proper preference relationships. Even so, we give a small numerical illustration, with a few stages of the construction (Fig. 12.8):

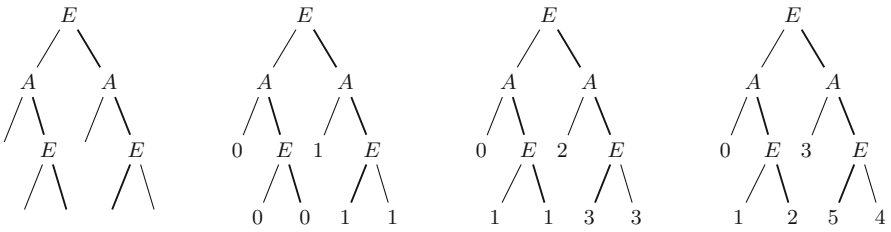


Fig. 12.8 Rationalizing top-down

12.6.2 Rationalization Procedures and Game Change

Now we are at a point where we can bring our preference dynamics of Chapter 4 to bear, in the same style as we did with information dynamics in earlier sections. In our view, a rationalization procedure is a natural game-theoretic process, just as much as Backward Induction itself, and it, too, invites logical analysis. But this time, the driving dynamics is not informational actions, but actions of *preference upgrade*.

We give an "on-line version" here in the earlier sense of following the actual top-down direction of play of a game. Our solution involves a slight generalization of the upgrades in Chapter 4, as will be clear from the proof of the following result:

Theorem 12.16 *Rationalization may be seen as successive preference upgrades following the observation of moves of a game.*

Proof We will not give a complete technical analysis here, but outline the main idea to a point where the reader can see how it works.

We can view the public observation of each successive move of a game according to the players' given strategies as a process with two steps, a *public announcement* followed by a *preference upgrade*:

[21] We leave the simple proof of correctness to the reader here.

- The first step is that we restrict, with a *PAL*-style update $!\varphi$, the space of all possible histories of the game to just those on which the given move occurs. Here the relevant assertion φ = "Move a was played", where a is the actual move.
- The next step "interprets" this observation to an upgrade $\Uparrow(\varphi)$ for the same assertion φ, at least when E is the active player. For, if the player that moves is E, then the histories following her chosen move must be (*made*) better for her than those following the other available moves that she did not take.

Iterating these steps will gradually introduce refined preferences in step with our earlier top-down algorithm. By the end of the game, the given strategies have become aligned with Backward Induction. This may again be proved by an induction, whose details we leave to the reader.

Actually, there is a little lack of precision here hat we must address. Consider any finite game tree. The description we just gave works at the first stage. But now consider the second stage of the game. The moves prescribed there can be different depending on the subgame that we are in. Thus, the upgrades that we talked about must be more complex, involving upgrades with assertions that are *relativized* to subsets of the total space of histories. To deal with this, we need new operations

$$\Uparrow^{\psi}(\varphi)$$

which make the φ-worlds better *only among the* ψ*-worlds* of the total model. One can define these formally, since their intended update effect is obvious.[22] One can then also add them to our earlier dynamic languages and write recursion axioms for them. We leave these details unspecified, since our main concern here is just making the connection between upgrade and rationalization. □

This "on-line" analysis applies to the above top-down algorithm. It would be of interest to also reconstruct the bottom-up algorithm, more in the "off-line" style of Section 12.4 with just *one single* action-preference upgrade that gets iterated until the right preference order is constructed. We have not been able to find such an assertion, so we must leave this as an open problem.

12.6.2.1 Dynamic Preference Logics of Game Change

We have hopefully said enough to make it plausible that, in this setting, our earlier dynamic preference logics directly apply to games, just as we saw earlier with public announcement logic for changing games:

Theorem 12.17 *The dynamic logic of information plus preference change in extensive games is completely axiomatizable.*

Proof The axiomatization merely puts together various components found in this book. On top of the chosen static base logic, we need the axioms for preference,

[22] For instance, it fits in the *PDL-program format* of Chapter 4.

action, and knowledge modalities under playing moves and public announcements of Chapters 4, 5, and Theorem 12.13. To these, we then merely add the reduction axioms for these same operators under radical upgrade as in Chapter 4.[23] □

12.6.2.2 Excursion: Game Change Under Entangled Dynamics

The preceding analysis fixed *BI* as a target of rationalization, given some freedom in construing players' preferences. But there are many other scenarios for reconstructing games. One would entangle preference with belief, the way we have done so often before. Here is a simple example.

Example 12.18 (rationalizing in terms of beliefs) Suppose A moves right in the game of our initial Example 12.1. One can interpret this rationally if we assume that *A believes that E will go right as well* in the next move. This rationalization is not in terms of preferences, these are now assumed to be given beforehand, but in terms of A's beliefs about E. Note that this style of rationalizing need not produce the *BI* solution – if the players' beliefs are too strange.

This, too, leads to rationalization procedures, as analyzed again in [28]. Consider a game where a strategy σ_A is given, as well as A's preferences (those of E do not matter in what follows). Assuming some minimal rationality again,[24] we (or the other player E) can rationalize this by assigning a suitable strategy to player E. Equivalently, as we have seen in an earlier section, this may be seen as assigning a suitable belief for A.

Example 12.19 (rationalization in beliefs) Figure 12.9 depicts a game with A's moves marked as bold-face arrows, and the necessary rationalizing beliefs indicated by the dotted arrows. The numbers at leaves stand for values for A.

In the diagram, A's initial choice for going left has been rationalized by forcing its outcome 4, by assuming that E will go left, which is better than the forced

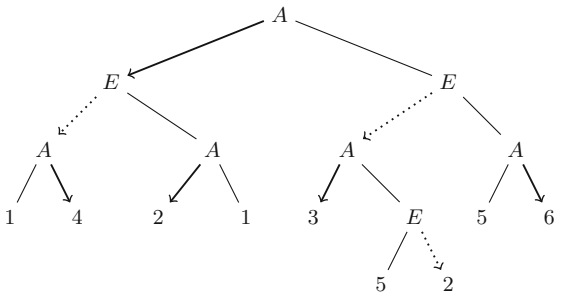


Fig. 12.9 Rationalization in beliefs

[23] The axiom for the move modality $\langle a \rangle$ is a simple operator commutation. A full version would need a small extension to "relativized upgrades" mentioned above.

[24] This time, the given σ_A never prescribes a move that is strictly dominated assuming that further play proceeds via the given strategy for A and *any moves* for E.

outcome 3 on the right – now assuming that E will go left there, too. One step further down, things are correct automatically for A's final choices, since minimal rationality guarantees that these are BI-optimal already. The only different case is the subtree with outcomes 3, 5, 2. Here the same procedure works as the initial one: We compare the outcome 3 with the outcome 2, and "force" the latter by the expectation that after the move not chosen, E would go right.

Again in what follows we make the argument in [28] a bit more precise. The general procedure is simply this:

Rationalizing in beliefs Working top-down, at turns for A, we consider his chosen move a. The minimal rationality tells us that for each alternative move b, there exists an endpoint u following a via further play along σ_A and any moves for E, and likewise such a reachable endpoint v after b such that A weakly prefers u to v.[25] Doing this for all alternative moves b, we can take the endpoint u with the *maximal value* for A here, which will work for all different v's chosen for different b. Now choose A's beliefs simply as follows:

> Following move a, A expects E to follow the moves leading to the endpoint u (given his playing σ_A), and following the other moves b, he expects E to take the moves leading to their endpoints v.

It is easy to see that, proceeding down the game tree, this belief assignment makes A's behavior rational: No move of his is ever "dominated-in-beliefs", in the sense of our earlier analysis of game solution.

This procedure of assigning beliefs can be analyzed in the plausibility-changing style that we have used in Section 12.4 for Backward Induction. We will not pursue this line here. Once again, the procedure may result in weird behavior, if we cannot find an independent reason for the postulated beliefs.[26]

12.6.2.3 Conclusion: Many Scenarios

With these examples, we have shown how games support many dynamic scenarios of reasoning, that involve both information change and preference change. Of course, the few examples that we looked at are just a start. For instance, it is not clear at all why we would want the rationalization to always end in the Backward Induction solution. If we drop the "minimal rationality" assumptions of the above algorithms, we still get outcomes, where players now have wilder styles of play, that can still make sense. Also, rationalization is just one procedure, and it is "off-line" in our earlier sense. Thus, it is not yet the real thing in terms of logics for "on-line" actual mind changes of players.

---

[25] Note the analogy with the Confluence Property pictured in Section 12.4.

[26] Reference [31] explains A's "irrational move" of going right in our running example in terms of running risks for the common good, expecting E to return the favor.

Even so, we have shown that preference dynamics and dynamic preference logics are a natural fit with natural issues concerning games, extending earlier work on their logical analysis.

12.6.3 Adding Priority to Game Representation

Next, we briefly consider the other main theme of this book: richer representations of preference including prioritized criteria that create betterness among worlds or outcomes of a game.

The notion of priority-based preference makes sense in games, too. Though we cannot read priorities directly from games, they are present in the form of underlying *goals* that each player wants to achieve.[27] Goals are expressed by propositions, ordered linearly or just partially, depending on the mental tidiness of players. From ordered goals, we can derive concrete preferences that players have over outcomes of the game, using the methods of Chapters 7 and 10.

What this additional structure allows us to do is make much finer comparisons between preferences of players, and the degree to which these are aligned.[28] Let us just illustrate two extreme cases in the setting of linear priority sequences:

(1) Two agents A and E are *cooperative* if they share the same priorities, ordered in the same manner. I.e., we have $D_1 \gg_a D_2 \gg \cdots \gg_a D_m (m \in \mathbb{N})$ and $D_1 \gg_b D_2 \gg \cdots \gg_b D_m (m \in \mathbb{N})$.
(2) Two agents are *competitive* if one has a priority order $D_1 \gg_a D_2 \gg \cdots \gg_a D_m (m \in \mathbb{N})$, while the other has $\neg D_m \gg_b \neg D_{m-1} \gg \cdots \gg_b \neg D_1 (m \in \mathbb{N})$.

In fact, Chapter 8 had several relevant completeness results for this setting:

Theorem 12.20 (cooperative and competitive agents) $\vdash_{\mathbf{KD45-PG}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences shared by two cooperative agents. Likewise,* $\vdash_{\mathbf{KD45-PG}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences for competitive agents.*

Clearly, most actual games will have players whose goals lie in between these two extremes.

Evaluating games in terms of priority But then, we can also define more fine-grained notions of equilibrium and purpose of a game, in terms of players trying to achieve shared goals, and competing on ones where they differ–guided by their priority structure. We will not explore this in detail, but here are some possible directions.

First, priority graphs suggest a generalization in games from connected to arbitrary *pre-orders* over outcomes. The main novelty is then that players can be either "indifferent" between outcomes, but also may find them "incomparable". Game

---

[27] A concrete *DEL*-related use of goals is made in the "knowledge games" of [1].

[28] Alignment can be very hard to judge if we just have players' extensional comparison lists of outcomes: just as we cannot say much about voters' similarities or dissimilarities in thinking if we only look at their voting records.

solution methods like Backward Induction can be adapted to this setting, since we can still define their driving notion of rationality like before, as avoiding strictly dominated moves. But clearly, this generalization changes players' evaluation of outcomes from "the best" to "some best". A yet more refined view of outcomes would retain the structure of priority graphs, and see which *subgraphs* players can realize through their strategies. In particular, I might "win", realizing some of my topmost goals, but even though you "lose" in one sense, you might still realize a considerable part of your prioritized agenda.

Another interesting use of priority graphs is the following. Chapter 10 extensively explored natural operations of graph merge, depending on how we construct authority relations between the players: either on a par (intersection of the individual preferences) or in a hierarchy (sequential composition of individual preferences). This offers a more refined view of possible *coalitions* in a game than the usual unstructured sets of players.

Finally, one more exciting viewpoint would be the one that we explored in Chapter 11 for deontics: How to use priorities for studying the *dynamics of play* in games. Many similar considerations apply. The set of goals might now change in the course of a game, and so might their alignment, changing the degree of "coordination" between the players. This, too, must remain on our list of things to do, though we do think that our illustrations in Chapter 11 might easily be extended to meaningful game-theoretic scenarios.

We will not pursue these suggestions here, but we do think they might lead us to rethink the usual numerical notions and methods for game solution.

This concludes our tentative exploration of how the main perspectives from this book can be linked to issues in the logical analysis of games.

12.7 Preference in a Long-Term Perspective

Finally, the finite games that we have considered are just one instance of something more general: the temporal evolution of information and evaluational processes. As we have observed in Chapter 2, one recent trend in dynamic epistemic logic has been the incorporation of crucial "procedural information" about the long term process one is in. Only such processes give direction to information gain and evaluation change. This study has resulted in merges of dynamic logics for local steps of information change with epistemic and doxastic temporal logics (cf. [35, 64]).

Very much the same is true for the preference dynamics of this book. Preferences tend to guide our behavior over time, and the latter may have structures of its own that are not reducible to single local steps of information or command. In fact, we have already seen some specific longer-term preference dynamics in this chapter, namely, in our rationalization procedures of Section 12.6. There, we considered what would happen in the limit of very simple procedures of repeatedly performing the same step of preference change.[29]

[29] Another relevant case would be the infinite games of *evolutionary game theory*: cf. [108].

Even in such simple scenarios, many logical subtleties emerge. For instance, it is not clear whether our procedures will always stabilize to one fixed model. Indeed, for the closely related case of plausibility change, [19] have shown that "cycles" can occur where repeated radical upgrade with the same assertion can lead to oscillating preference patterns. Thus, the *temporal logic* of preference change might be full of surprises. On the positive side, though, there are also many instances where preference logic might use already existing results. In particular, we think that the complete and decidable "protocol versions" of dynamic-epistemic logic found in [110] can easily be generalized to deal with preference change. Likewise, we think that the methods of this book could combine with the logics of explicit protocols in [202].

Clearly, then, the story of this chapter is just the start of a longer road ahead.

12.8 Conclusion

We have shown how preference logic fits naturally with games, the prime area where information and evaluation meet. We have seen in some detail how the technical notions of this book occur there, and we have suggested how our new themes of dynamics and priority might have a contribution to make to the currently emerging interfaces of logic and game theory.

Part VII
Finale

Chapter 13
Conclusion

This book has presented a uniform logical theory of preference, drawing together ideas from several areas: modal logics of betterness relations, dynamic epistemic logics of information change, and priority-based systems of representing structured relations. Our chapters have added successive components of this theory, showing how it can be developed with logical techniques, suitably adapted to deal with preference structure. The result is a framework that has interesting theoretical features of its own, witness the sequence of technical results in this book on completeness, definability, and other architectural features of our systems. But we can also use the system of this book for analyzing preferences in a wide variety of fields, from epistemology and ethics to computer science and game theory. The evidence for this so far consists in a few case studies on deontic reasoning and on games in our final chapters, while we would also mention the use of our techniques in modeling beliefs and belief revision, thanks to the analogy between betterness and plausibility relations.

I would also like to summarize the Grand Picture that has driven this investigation for me. When I look at human behavior, what strikes me most is the duality of two systems in all of language use and agency: the dynamics of *information* and of *evaluation*. It seems to me that only that interplay "makes sense" of what we do in the full meaning of that phrase, and only their harmony creates truly "rational" behavior. This book has been an attempt at providing a clearer picture of how these two crucial systems show analogous static and dynamic logical structure, and also, how they can live in harmony, entangled in many ways that we do not yet always understand fully. That I see as the real message behind the array of technical logics with epistemic, doxastic, and preference features developed in this book.

I also hope to have convinced the reader that, in doing all this, preference logic can be "much richer than it is", including a broader view of representation than simple modal structures, and taking preference dynamics on board as a serious and essential part of the enterprize. Moreover, I hope to have shown how the framework of this book, once mastered, can yield new applications beyond the concrete pilot examples that we have given. But that, of course, also depends on what further studies will be undertaken from the perspective offered here. I conclude with a few topics that are on my own agenda right now.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4_13,

Of course, along the way, this book has left many loose ends, with open problems about exploring the systems that we have defined. But more than that, I would mention four further directions.

First, I have mostly considered preferences for single agents, while rational agency clearly also involves *groups*. We would like to study social preferences, which are tied up with social relations in communities. For instance, there is a distinction between my preferences, my friends' preferences and our aggregated preferences. A first exploration of a logic of communities may be found in [169]. A similar distinction plays in social choice theory once we start analyzing the deliberation phase prior to voting where individual preferences may shift, while collective ones get created. Technically, it is a simple task to extend everything in this book to a multi-agent setting – and the main reason why we have not done so was a desire to avoid prolixity of subscript notation. But thinking about what would be collective preferences, or even simple counterparts to the usual epistemic notions of common and distributed knowledge or belief for groups seems much harder. The priority graphs of [6] that we used in Chapter 10 for representing reason-based preference of a single agent, were originally intended to model preference merge for groups of agents. This seems a good starting point, but the real work remains to be done.

Next, it should be noted that, just like information dynamics, preference change does not just involve myopic single steps. It often takes place in *longer scenarios over time*, when agents try to achieve goals in the long run. Of course, we can iterate the dynamic operators in our logics to describe specified numbers of consecutive "local dynamic" updates or upgrades. But we also want limit behavior without a preset bound. To some extent, this theme has surfaced in Chapter 12 on preference in extensive games that model strategic behavior of agents interacting over time. To fully understand the logical temporal dynamics of preference, we need to integrate time into the current framework, as has been done for dynamic epistemic and doxastic logic in [45, 64, 158, 201], and other publications. For a start, we need to see what preference adds here in terms of features of its own.

While both preceding topics might be seen as moving to larger scales: in aggregation of agents, or in number of time steps, there is also an open problem of smaller scale and *more fine-grained syntactic representation*. The treatment of preference in this book has been mostly at the standard semantic level of semantic information, suppressing distinctions between logically equivalent expressions. While this is sound methodology, it also leaves out some crucial things. In the area of information dynamics, one of these things is the dynamic role of inference: obviously significant, but actually, producing no change at all in current semantic ranges. There are some attempts at creating more fine-grained syntactic dynamic logics of acts of inference and reasoning (cf. [191]), but no general framework has emerged yet. Likewise, the theme of syntax has been an undercurrent in this book. One part of this is the same as for information dynamics: Acts of inference can affect our preferences as much as our knowledge or belief. Once I become explicitly aware of some features of an object, my preference for it may change. And also in our priority graphs for structured preferences, we saw syntactic structure right up front: These graphs are syntactic objects, admitting of syntactic manipulation. To make it very concrete:

A "change in the law" tends to be a syntactic change, not a semantic one. I see a systematic development of more fine-grained syntactic aspects of preference as a third major desideratum.

Finally, there are also evident broader questions relating our logical framework to other approaches. In particular, I have adopted a qualitative approach to preference representation. But in areas like decision theory, game theory, and social choice theory, usually, numerical utility functions represent preference. Likewise, for modeling beliefs under uncertainties, numerical probabilities are used widely. In the area of belief revision, and to some extent also in *DEL* these days (cf. [10, 36, 165], and [15]), this is a well-known interface. Can the logical systems for preference proposed in this book interface with *quantitative* utilities in a natural manner? I appended a first attempt in Chapter 6, using *DEL* methodology to upgrade numerical "plausibility values", using ideas from [10] and [130], but much more remains to be done.

I am not saying that it will be easy to fulfill all of these further desiderata. But I do think that the logical perspective offered in this book has opened up a way to go.

References

1. Ågotnes, T., and H. van Ditmarsch. 2011. What will they say? – public announcement games. To appear in *Synthese* 179:57–85.
2. Alchourrón, C., P. Gärdenfors, and D. Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50:510–530.
3. Alechina, N., M. Jago, and B. Logan. 2007. Belief revision for rule-based agents. In *A meeting of the minds–proceedings of the workshop on logic, rationality and interaction*, eds. J. van Benthem, S. Ju, and F. Veltman, 99–112. London: King's College Publications.
4. Alechina, N., B. Logan, and M. Whitsey. 2004. A complete and decidable logic for resource-bounded agents. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2004),* eds. N.R. Jennings, C. Sierra, L. Sonenberg, and M. Tambe, 606–613, ACM Press: New York.
5. A.R. Anderson. 1957. The formal analysis of normative concepts. *American Sociological Review*, 22:9–17.
6. Andréka, H., M. Ryan, and P-Y. Schobbens. 2002. Operators and laws for combining preferential relations. *Journal of Logic and Computation* 12:12–53.
7. Åqvist, L. 1967. Good Samaritans, contrary-to-duty imperatives and epistemic obligations. *Nous* 1:361–379.
8. Åqvist, L. 1994. Deontic logic. In *Handbook of philosophical logic*, eds. D. Gabbay and F. Guenthner, volume 2, 605–714. Dordrecht: Kluwer.
9. Areces, C., and B. ten Cate. 2006. Hybrid logics. In *Handbook of modal logic*, eds. P. Blackburn, J. van Benthem, and F. Wolter, 821–868. Amsterdam: Elsevier.
10. Aucher, G. 2003. A combined system for update logic and belief revision. Master's thesis, MoL-2003-03. ILLC, University of Amsterdam.
11. Aucher, G. 2008. Perspectives on belief and change. PhD thesis, Université de Toulouse.
12. Aucher, G., D. Grossi, A. Herzig, and E. Lorini. 2009. Dynamic context logic. In *Proceedings of the 2nd international workshop on logic, rationality and interaction (LORI 2009)*, eds. X. He, J. Horty, and E. Pacuit, volume 5834 of *FoLLI-LNAI*, 15–26. Springer: Heidelberg.
13. Balbiani, P., V. Goranko, R. Kellerman, and D. Vakarelov. 2007. Logical theories for fragments of elementary geometry. In *Handbook of spatial logics*, eds. M. Aiello, I. Pratt-Hartmann, and J. van Benthem, 343–428. Springer: Dordrecht.
14. Baltag, A., L.S. Moss, and S. Solecki. 1998. The logic of common knowledge, public announcements, and private suspicions. In *Proceedings of the 7th conference on theoretical aspects of rationality and knowledge (TARK 98)*, ed. I. Gilboa, 43–56. Morgan Kaufmann Publishers: San Francisco, CA, USA.
15. Baltag, A., and S. Smets. 2006. 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.
16. Baltag, A., and S. Smets. 2006. The logic of conditional doxastic actions: A theory of dynamic multi-agent belief revision. In *Proceedings of the workshop on rationality and knowledge*. eds. S. Artemov and R. Parikh, 13–30. Malaga: ESSLLI.

F. Liu, *Reasoning about Preference Dynamics*, Synthese Library 354,
DOI 10.1007/978-94-007-1344-4, © Springer Science+Business Media B.V. 2011

17. Baltag, A., and S. Smets. 2008. Probabilistic dynamic belief revision. *Synthese* 165(2): 179–202.
18. Baltag, A., and S. Smets. 2008. A qualitative theory of dynamic interactive belief revision. In *Logic and the foundations of game and decision theory*, eds. M. Wooldridge, G. Bonanno, and W. van der Hoek, volume 3 of *Texts in Logic and Games*, 9–58. Amsterdam: Amsterdam University Press.
19. Baltag, A., and S. Smets. 2009. Group belief dynamics under iterated revision: Fixed points and cycles of joint upgrades. In *Proceedings of the 12th conference on theoretical aspects of rationality and knowledge (TARK 2009)*, ed. A. Heifetz, 41–50. Morgan Kaufmann Publishers: San Francisco, CA, USA.
20. Baltag, A., S. Smets, and J. Zvesper. 2009. Keep 'hoping' for rationality: A solution to the backward induction paradox. *Synthese* 169(2):301–333.
21. Barbera, S., W. Rossert, and P.K. Prasanta. 2001. Ranking sets of objects. Departement de sciences economiques, Universite de Montreal. See http://www.sceco.umontreal.ca/publications/etext/2001-02.pdf.
22. van Benthem, J. 1982. Later than late: On the logical origin of the temporal order. *Pacific Philosophical Quarterly* 63:193–203.
23. van Benthem, J. 1996. *Exploring logical dynamics*. Stanford: CSLI Publications.
24. van Benthem, J. 1999. Modal correspondence theory. In *Handbook of philosophical logic*, eds. D. Gabbay and F. Guenthner, volume 3, 325–408. Dordrecht: Kluwer, second edition. Reprint with addenda.
25. van Benthem, J. 2001. Games in dynamic-epistemic logic. *Bulletin of Economic Research* 53:219–248.
26. van Benthem, J. 2005. Open problems in logic and games. In *Essays in honour of Dov Gabbay*, eds. S. Artemov et al., 229–264. London: King's College Publications.
27. van Benthem, J. 2006. 'One is a lonely number': On the logic of communication. In *logic colloquium*, eds. P. Koepke Z. Chatzidakis, and W. Pohlers, *ASL Lecture Notes in Logic 27*. Providence, 96–129. RI: AMS Publications.
28. van Benthem, J. 2006. Rationalizations and promises in games. In *Philosophical trend: Special issue in logic*, 1–6. Beijing: The Chinese Association of Logic.
29. van Benthem, J. 2007. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logic* 17:129–156.
30. van Benthem, J. 2007. In praise of strategies. In *Foundations of social software*, eds. J. van Eijck and R. Verbrugge, 283–317. London: King's College Publications.
31. van Benthem, J. 2007. Rational dynamics and epistemic logic in games. *International Game Theory Review* 9:13–45.
32. van Benthem, J. 2011. *Logical dynamics of information and interaction*. Cambridge: Cambridge University Press.
33. van Benthem, J., J. van Eijck, and A. Frolova. 1993. Changing preferences. Technical Report, CS-93-10, Centre for Mathematics and Computer Science, Amsterdam.
34. van Benthem, J., J. van Eijck, and B. Kooi. 2006. Logics of communication and change. *Information and Computation* 204:1620–1662.
35. van Benthem, J., J. Gerbrandy, T. Hoshi, and E. Pacuit. 2009. Merging frameworks for interaction. *Journal of Philosophical Logic* 38(5):491–526.
36. van Benthem, J., J. Gerbrandy, and B. Kooi. 2006. Dynamic update with probabilities. Technical Report, PP-2006-21, ILLC, University of Amsterdam.
37. van Benthem, J., J. Gerbrandy, and B. Kooi. 2009. Dynamic update with probabilities. *Studia Logica* 93(1):67–96.
38. van Benthem, J., and A. Gheerbrant. 2010. Epistemic dynamics and fixed-point logics. *Fundament Informatica* 100(1–4):19–41.
39. van Benthem, J., P. Girard, and O. Roy. 2009. Everything else being qqual: A modal logic approach for ceteris paribus preferences. *Journal of Philosophical Logic* 38(1):83–125.

40. van Benthem, J., and D. Grossi. 2011. Normal forms for priority graphs. Technical Report PP-2011-02, ILLC, University of Amsterdam.
41. van Benthem, J., and F. Liu. 2004. Diversity of logical agents in games. *Philosophia Scientiae* 8:163–178.
42. van Benthem, J., and F. Liu. 2007. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logic* 17:157–182.
43. van Benthem, J., and S. Minica. 2009. Toward a dynamic logic of questions. In *Proceedings of the 2nd international workshop on logic, rationality and interaction (LORI 2009)*, eds. X. He, J. Horty, and E. Pacuit, volume 5834 of *FoLLI-LNAI*, 27–41. Springer: Heidelberg.
44. van Benthem, J., S. van Otterloo, and O. Roy. 2006. Preference logic, conditionals and solution concepts in games. In *Modality matters: Twenty-five essays in honour of Krister Segerberg*, eds. H. Lagerlund, S. Lindström, and R. Sliwinski, 61–77. Uppsala Philosophical Studies 53. Uppsala University: Uppsala.
45. van Benthem, J., and E. Pacuit. 2006. The tree of knowledge in action: Towards a common perspective. In *Proceedings of advances in modal logic (AiML 2006)*, eds. G. Governatori, I. Hodkinson, and Y. Venema, 87–106. London: King's College Publications.
46. van Benthem, J., E. Pacuit, and O. Roy. 2011. From game theory to theory of play, a logical perspective. To appear In *Games* ed. H. Gintis Special Issue on Epistemic Game Theory and Modal Logic. *Games* 2:52–86.
47. van Benthem, J. 2010. *Modal logic for open minds*. Stanford: CSLI Publications.
48. van Benthem, J., D. Grossi, and F. Liu. 2010. Deontics = betterness + priority. In *Deontic logic in computer science, 10th international conference, DEON 2010*, eds. G. Governatori and G. Sartor, volume 6181 of *LNAI*, 50–65. Springer: Heidelberg.
49. Blackburn, P. 2000. Representation, reasoning, and relational structures: A hybrid logic manifesto. *Logic Journal of the IGPL* 8:339–365.
50. Blackburn, P., M. de Rijke, and Y. Venema. 2001. *Modal logic*. Cambridge: Cambridge University Press.
51. Boella, G., G. Pigozzi, and L. van der Torre. 2009. Normative framework for normative system change. In *Proceedings of the eighth international conference on autonomous agents and multiagent systems (AAMAS 2009)*, eds. P. Decker, J. Sichman, C. Sierra, and C. Castelfranchi, 169–176. IFAAMAS Publications.
52. Bolle, F. 1983. On Sen's second-order preferences, morals, and decision theory. *Erkenntnis* 20:195–205.
53. Boutilier, C. 1990. Conditional logics of normality as modal systems. In *Proceedings of AAAI'90*, 594–599. Boston, MA.
54. Boutilier, C. 1993. A modal characterization of defeasible deontic conditionals and conditional goals. In *AAAI spring symposium on reasoning about mental states: Formal theories and applications*, 30–39, Stanford.
55. Boutilier, C. 1994. Conditional logics of normality: A modal approach. *Artificial Intelligence* 68:87–154.
56. de Bruin, B. 2004. *Explaining games: On the logic of game theoretic explanations*. PhD thesis, ILLC, University of Amsterdam.
57. Chevaleyre, Y., U. Endriss, and J. Lang. 2006. Expressive power of weighted propositional formulas for cardinal preference modeling. In *Proceedings of the 10th international conference on principles of knowledge representation and reasoning (KR 2006)*, eds. P. Doherty, J. Mylopoulos, and C. Welty, 145–152. Menlo Park, CA: AAAI Press.
58. Chisholm, R. 1982. *The foundations of knowing*. Minneapolis, MN: University of Minnesota Press.
59. Chisholm, R., and E. Sosa. 1966. Intrinsic preferability and the problem of supererogation. *Synthese* 16:321–331.
60. Cohen, P.R., and H.J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence* 42:213–361.

61. Conee, E., and R. Feldman. 2004. *Evidentialism. Essays in epistemology*. Oxford: Oxford University Press.
62. Coste-Marquis, S., J. Lang, P. Liberatore, and P. Marquis. 2004. Expressive power and succinctness of propositional languages for preference representation. In *Proceedings of the 9th international conference on principles of knowledge representation and reasoning (KR 2004)*, eds. D. Dubois, C. Welty, and M.-A. Williams, 203–212. Menlo Park, CA: AAAI Press.
63. Coulhon, T., and P. Mongin. 1989. Social choice theory in the case of von neumann-morgenstern utilities. *Soc Choice Welfare* 6:175–187.
64. Dégremont, C. 2010. The temporal mind. Observations on the logic of belief change in interactive systems. PhD thesis, ILLC, University of Amsterdam.
65. van Ditmarsch, H., W. van der Hoek, and B. Kooi. 2007. *Dynamic epistemic logic*. Berlin: Springer.
66. Doyle, J., Y. Shoham, and M.P. Wellman. 1991. A logic of relative desire. In Methodologies for Intelligent Systems, eds. Z.W. Ras and M. Zemankova, Springer: Berlin.
67. Doyle, J., and R.H. Thomason. 1999. Background to qualitative decision theory. *AI Magazine* 20:55–68.
68. Doyle, J., and M.P. Wellman. 1994. Representing preferences as Ceteris Paribus comparatives. In *Proceedings of the AAAI spring symposium on decision-theoretic planning*, 69–75. Stanford.
69. Dretske, F. 1981. *Knowledge and the flow of information*. Oxford: Blackwell.
70. Ebbinghaus, H-D., and J. Flum. 1995. *Finite model theory. Perspectives in mathematical logic*. Berlin: Springer.
71. van Eck, J. 1982. A system of temporally relative modal and deontic predicate logic and its philosophical applications. PhD thesis, University of Amsterdam.
72. Elster, J. 1983. *Sour Grapes. Studies in the subversion of rationality*. Cambridge: Cambridge University Press.
73. Fagin, R., J.Y. Halpern, Y. Moses, and M.Y. Vardi. 1995. *Reasoning about knowledge*. Cambridge, MA: The MIT Press.
74. Fishburn, P.C. 1970. *Utility theory for decision making*. Huntington, NY: Robert E. Krieger Publishing Company.
75. Fishburn, P.C. 1973. *The theory of social choice*. Princeton, NJ: Princeton University Press.
76. Fishburn, P.C. 1999. Preference structures and their numerical representations. *Theoretical Computer Science* 217:359–383.
77. Forrester, J. 1984. Gentle murder, or the adverbial samaritan. *The Journal of Philosophy* 81: 193–197.
78. van Fraassen, B. 1973. Values and the heart's command. *The Journal of Philosophy* 70(1): 5–19.
79. Gargov, G., and V. Goranko. 1993. Modal logic with names. *Journal of Philosophical Logic* 22(6):607–636.
80. Garson, J. 2001. Quantification in modal logic. In *Handbook of Philosophical Logic,* eds. D. Gabbay and F. Guenthner, second edition, volume 3, Dordrecht: D. Reidel, 267–323.
81. Gerbrandy, J. 1999. Bisimulation on planet Kripke. PhD thesis, ILLC, University of Amsterdam.
82. Gheerbrant, A. 2010. *Fixed-point logics on trees*. PhD thesis, ILLC, University of Amsterdam.
83. Girard, P. 2008. *Modal logics for belief and preference change*. PhD thesis, Stanford University.
84. Goldblatt, R.I. 1982. *Axiomatizing the logic of computer programming*, volume 130 of *LNCS*. Springer: Berlin.
85. Goldman, A. 1991. *Epistemic folkways and scientific epistemology*. Liaisons: Philosophy Meets the Cognitive and Social Sciences. Cambridge, MA: The MIT Press.
86. Governatori, G., and A. Rotolo. 2005. Logic of violations: A gentzen system for reasoning with contrary-to-duty obligations. *Australasian Journal of Logic* 3:193–215.

87. Governatori, G., and A. Rotolo. 2008. Changing legal systems: Abrogation and annulment. part 1: Revision and defeasible theories. In *Proceedings of the 9th international conference on deontic logic in computer science (DEON 2008)*, eds. R. van der Meyden and L. van der Torre, volume 5076 of *LNAI*, 3–18. Springer: Berlin.
88. Governatori, G., and A. Rotolo. 2008. Changing legal systems: Abrogation and annulment. part 2: Temporalised defeasible logic. In *Proceedings of the 3rd international workshop on normative multiagent system (NorMAS 2008), eds. G. Boella, G. Pigozzi, M. P. Singh, and H. Verhagen, Luxembourg, Luxembourg, July 14–15, 2008*, 112–127.
89. Grosof, B.N. 1991. Generalising prioritization. In *Proceedings of the 2th international conference on principles of knowledge representation and reasoning (KR 91)*, eds. J. Allen, R. Fikes, and E. Sandewall, 289–300. San Fransisco, CA: Morgan Kaufmann.
90. Grossi, D. and F. Velazquez-Quesada. 2009. Twelve angry men: A study on the fine-grain of announcements. In *Proceedings of the 2nd international workshop on logic, rationality and interaction (LORI 2009)*, eds. X. He, J. Horty, and E. Pacuit, volume 5834 of *FoLLI-LNAI*, 147–170. Springer: Heidelberg.
91. Grove, A. 1988. Two modelings for theory change. *Journal of Philosophical Logic* 17: 157–170.
92. Grune-Yanoff, T., and S.O. Hansson. eds. 2009. *Preference change: Approaches from philosophy, economics and psychology*. Theory and Decision Library. Springer: Heidelberg.
93. Halldén, S. 1957. *On the logic of "better"*. CWK Gleerup: Lund.
94. Halpern, J.Y. 1997. Defining relative likelihood in partially-ordered preferential structure. *Journal of Artificial Intelligence Research* 7:1–24.
95. Hansson, B. 1968. Fundamental axioms for preference relations. *Synthese* 18:423–442.
96. Hansson, B. 1969. An analysis of some deontic logics. *Nous* 3:373–398.
97. Hansson, S.O. 1990. Defining 'good' and 'bad' in terms of 'better'. *Notre Dame of Journal of Formal Logic* 31:136–149.
98. Hansson, S.O. 1990. Preference-based deontic logic. *Journal of Philosophical Logic* 19: 75–93.
99. Hansson, S.O. 1992. Similarity semantics and minimal changes of belief. *Erkenntnis* 37: 401–429.
100. Hansson, S.O. 1995. Changes in preference. *Theory and Decision* 38:1–28.
101. Hansson, S.O. 2001. Preference logic. In *Handbook of philosophical Logic*, eds. D. Gabbay and F. Guenthner, volume 4, 319–393. Dordrecht: Kluwer.
102. Hansson, S.O. 2001. *The structure of values and norms*. Cambridge: Cambridge University Press.
103. Hansson, S.O. and T. Grüne-Yanoff. 2006. Preferences. In *Stanford encyclopedia of philosophy*. ed. E.N. Zalta, Stanford. http://plato.stanford.edu/entries/preferences/.
104. Harel, D., D. Kozen, and J. Tiuryn. 2000. *Dynamic logic*. Cambridge, MA: The MIT Press.
105. Harrenstein, P. 2004. *Logic in conflict. Logical explorations in strategic equilibrium*. PhD thesis, Utrecht University.
106. van der Hoek, W., and M Pauly. 2006. Modal logic for games and information. In eds. J. van Benthem, F. Wolter, P. Blackburn, *Handbook of modal logic*, 1077–1148. Amsterdam: Elsevier.
107. van der Hoek, W., and M. Wooldridge. 2003. Towards a logic of rational agency. *Logic Journal of the IGPL* 11:133–157.
108. Hofbauer, J., and K. Sigmund. 1998. *Evolutionary games and population dynamics*. Cambridge: Cambridge University Press.
109. Holliday, W.H. 2009. Dynamic testimonial logic. In *Proceedings of the 2nd international workshop on logic, rationality, and interaction (LORI 2009)*, eds. X. He, J. Horty, and E. Pacuit, volume 5834 of *FoLLI-LNAI*, 161–179. Springer: Heidelberg.
110. Hoshi, T. 2009. *Epistemic dynamics and protocol information*. PhD thesis, Stanford University.
111. Houser, D., and R. Kurzban. 2002. Revealed preference, belief, and game theory. *Economics and Philosophy* 16:99–115.

112. Hughes, R. 1980. Rationality and intransitive preferences. *Analysis* 40:132–134.
113. Jeffrey, R.C. 1965. *The logic of decision*. Chicago, IL: University of Chicago Press.
114. Jeffrey, R.C. 1977. A note on the kinematics of preference. *Erkenntnis* 11:135–141.
115. Hinttika, J. 1962. *Knowledge and belief*. Ithaca, NY: Cornell University Press.
116. Jones, A.J.I., and I. Pörn. 1985. Ideality, sub-ideality and deontic logic. *Synthese* 275–290.
117. de Jongh, D., and F. Liu. 2009. Preference, priorities and belief. In eds. T. Grune-Yanoff and S.O. Hansson, *Preference change: Approaches from philosophy, economics and psychology*, 85–108. Theory and Decision Library. Springer: Heidelberg.
118. Ju, F. 2010. Imperatives and logic *Studies in Logic* 3(2):361–379.
119. Ju, F., and F. Liu. 2010. Adding priorities to update semantics for imperatives. Manuscript.
120. Kanger, S. 1971. New fondations for ethical theory. In ed. R. Hilpinen, *Deontic logic: Introductory and systematic readings*, 36–58. Dordrecht: D. Reidel.
121. Kratzer, A. 1981. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10:201–216.
122. Lang, J., and L. van der Torre. 2008. From belief change to preference change. In *Proceedings of the 18th European conference on artificial intelligence (ECAI-2008)*, 351–355. IOS Press: Amsterdam.
123. Lang, J., 2003. 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)*, 189–231. Morgan Kaufmann Publishers: San Francisco, CA, USA.
124. Lee, R. 1984. Preference and transitivity. *Analysis* 44:129–134.
125. Lenzen, W. 1983. On the representation of classificatory value structures. *Theory and Decision* 15:349–369.
126. Lewis, D. 1969. *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.
127. Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.
128. Lewis, D. 1981. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic* 10:217–234.
129. van Linder, B., W. van der Hoek, and J-J.Ch. Meyer. 1996. Formalising motivational attitudes of agents: On preferences, goals and commitments. In *Intelligent agents volume II – agent theories, architectures, and languages (ATAL '96)*, eds. M. Wooldridge, J. Mueller, and M. Tambe, 17–32. Berlin: Springer.
130. Liu, F. 2004. *Dynamic variations: Update and revision for diverse agents*. Master's thesis, MoL-2004-05. ILLC, University of Amsterdam.
131. Liu, F. 2008. *Changing for the better: Preference dynamics and agent diversity*. PhD thesis, ILLC, University of Amsterdam.
132. Liu, F. 2009. Diversity of agents and their interaction. *Journal of Logic, Language and Information* 18(1):23–53.
133. Liu, F. 2009. Preference change: A quantitative approach. *Studies in Logic* 2(3):12–27.
134. Liu, F. 2010. Von Wright's 'The Logic of Preference' revisited. *Synthese* 175(1):69–88.
135. Liu, F. 2011. A two-level perspective on preference. *Journal of Philosophical Logic* 40(3):421–439.
136. Loohius, L.O. 2009. Obligations in a responsible world. In *Proceedings of the 2nd international workshop on logic, rationality and interaction (LORI 2009)*, eds. X. He, J. Horty, and E. Pacuit, volume 5834 of *FoLLI-LNAI*, 251–262. Springer: Heidelberg.
137. Lorini, E., and F. Schwarzentruber. 2010. A modal logic of epistemic games. *Games* 1(4):478–526.
138. Lutz, C. 2006. Complexity and succinctness of public announcement logic. In *Proceedings of the 5th international conference on autonomous agents and multiagent systems (AAMAS06)*, 137–144. IEEE/ACM Press: New York.
139. Mastop, R. 2005. *What can you do? Imperative mood in semantic theory*. PhD thesis, ILLC, University of Amsterdam.

140. McCarthy, J. 1986. Circumscription – A form of non-monotonic reasoning. *Artificial Intelligence* 13:27–39.
141. van der Meyden, R. 1996. The dynamic logic of permission. *Journal of Logic and Computation* 6:465–479.
142. Meyer, J.-J.Ch. 1988. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic* 29:109–136.
143. Meyer, J.-J.Ch. and W. van der Hoek. 1995. *Epistemic logic for computer science and artificial intelligence*. Cambridge: Cambridge University Press.
144. Marx, M. 2006. Complexity of modal logic. In *Handbook of modal logic*, eds. J. van Benthem, F. Wolter, P. Blackburn, 139–180. Amsterdam: Elsevier.
145. Nayak, A. 1994. Iterated belief change based on epistemic entrenchment. *Erkentnis* 41: 353–390.
146. Nozick, R. 1981. *Philosophical explanations*. Cambridge, MA: Harvard University Press.
147. Osborne, M., and A. Rubinstein. 1994. *A course in game theory*. Cambridge, MA: The MIT Press.
148. van Otterloo, S., W. van der Hoek, and M. Wooldridge. 2006. Knowledge condition games. *Journal of Logic, Language, and Information* 15(4):425–452.
149. Pacuit, E., R. Parikh, and E. Cogan. 2006. The logic of knowledge based on obligation. *Synthese* 149:311–341.
150. Dekker, P. 2008. A guide to dynamic semantics. Technical Report PP-2008-42, ILLC, University of Amsterdam.
151. Plantinga, A. 1991. *Knowledge in perspective. Selected essays in epistemology*. Cambridge: Cambridge University Press.
152. Plaza, J.A. 1989. Logics of public communications. In *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, eds. M. Emrich, M. Pfeifer, M. Hadzikadic, Z. Ras, 201–216. Charlotte, North Carolina.
153. Prakken, H., and M. Sergot. 1996. Contrary-to-duty obligations. *Studia Logica* 57:91–115.
154. Prince, A., and P. Smolensky. 2004. *Optimality theory: Constraint interaction in generative grammar*. Oxford: Blackwell.
155. Rao, A.S., and M.P. Georgeff. 1991. Modeling rational agents within a BDI-architecture. In *Proceedings of the 2nd international conference on principles of knowledge representation and reasoning*, eds. J. Allen, R. Fikes, and E. Sandewall, 473–484. San Mateo, CA: Morgan Kaufmann.
156. Aumann, R. 1976. Agreeing to disagree. *Annals of Statistics* 4:1236–1239.
157. Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13(1–2):81–132.
158. Renne, B., J. Sack, and A. Yap. 2009. Dynamic epistemic temporal logic. In *Proceedings of the 2nd international workshop on logic, rationality and interaction (LORI 2009)*, eds. J. Horry X. He and E. Pacuit, volume 5834 of *FoLLI-LNAI*, 263–277. Springer: Heidelberg.
159. Rescher, N. 1966. Notes on preference, utility, and cost. *Synthese* 16:332–343.
160. Rooij, R. van. 2011. Revealed Preference and Satisficing Behavior, *Synthese* 179:1–12.
161. Rott, H. 2001. *Change, choice and inference: A study of belief and revision and nonmonotonic reasoning*. Oxford: Oxford University Press.
162. Rott, H. 2003. Basic entrenchment. *Studia Logica* 73:257–280.
163. Rott, H. 2006. Shifting priorities: Simple representations for 27 iterated theory change operators. In *Modality matters: Twenty-five essays in honour of Krister Segerberg*, eds. H. Langerlund, S. Lindström, and R. Sliwinski, 359–384. Uppsala Philosophical Studies 53. Uppsala University: Uppsala.
164. Roy, O. 2008. *Thinking before acting: Intentions, logic and rational choice*. PhD thesis, ILLC, University of Amsterdam.
165. Sack, J. 2009. Extending probabilistic dynamic epistemic logic. *Synthese* 169(2):241–257.
166. Savage, L.J. 1954. *The foundations of statistics*. New York, NY: Wiley.
167. Schumm, G. 1975. Remark on a logic of preference. *Notre Dame Journal of Formal Logic* 16:509–510.

168. Seligman, J. 2010. Hybrid logic for analyzing games. Working paper.
169. Seligman, J., F. Liu, and P. Girard. 2011. Logic in the community. In *Proceedings of the 4th Indian conference on logic and its applications*, eds. M. Banerjee and A. Seth, volume 6521 of *LNCS*, 178–188. Berlin: Springer.
170. Sen, A. 1971. Choice functions and revealed preference. *Review of Economic Studies* 38:307–317.
171. Sen, A. 1973. Behaviour and the concept of preference. *Economica* 40:241–259.
172. Sergot, 2004. $(C+)^{++}$: An action language for modeling norms and institutions. Technical Report 8, Department of Computing, Imperial College, London.
173. Shoham, Y. 1988. *Reasoning about change: Time and causation from the standpoint of artificial intelligence*. Cambridge, MA: The MIT Press.
174. Shoham, Y., and K. Leyton-Brown. 2008. *Multiagent systems: algorithmic, game theoretic and logical foundations*. Cambridge: Cambridge University Press.
175. Snyder, J. 2004. Product update for agents with bounded memory. Manuscript, Department of Philosophy, Stanford University.
176. Spohn, W. 1988. Ordinal conditional functions: A dynamic theory of epistemic states. In *Causation in decision, belief change and statistics II*, eds. W.L. Harper and B. Skyrms, 105–134. Dordrecht: Kluwer.
177. Paul, St. 1901. First letter to the Corinthians, Chapter 7, volume 9. Standard American Edition of the Bible.
178. Aquinas, St T. Summa Theologiae. 1951. Translation by Thomas Gilby, St. Thomas Aquinas: Philosophical Texts, Oxford University Press.
179. Stalnaker R. 1996. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy* 12(2):133–163.
180. Steup, M., and E. Sosa, ed. 2005. *Contemporary debates in epistemology*. Cambridge, MA: Blackwell.
181. Tan, Y.-H., and L. van der Torre. 1996. How to combine ordering and minimizing in a deontic logic based on preferences. In *Deontic logic, agency and normative systems, DEON '96: The 3rd international workshop on deontic logic in computer science*, eds. M. Brown and J. Carmo, 216–232. Springer: Heidelberg.
182. Thomason, R. 2000. Desires and defaults: A framework for planning with inferred goals. In *Proceedings of the 7th International Conference on Principles of Knowledge Representation and Reasoning (KR 2000)*, eds. A.G. Cohn, F. Giunchiglia, and B. Selman, 702–713. Morgan Kaufmann Publishers: San Francisco, CA, USA.
183. van der Torre, L. 1997. *Reasoning about obligations: Defeasibility in preference-based deontic logic*. PhD thesis, Rotterdam.
184. van der, L., Torre and Y. Tan. 1998. The temporal analysis of Chisholm's paradox. In *AAAI '98/IAAI '98: Proceedings of the 14th national/10th conference on artificial intelligence/innovative applications of artificial intelligence*, eds. T. Senator and B. Buchanan, 650–655. Menlo Park, CA: AAAI Press.
185. van der Torre, L. and Y. Tan. 1999. An update semantics for deontic reasoning. In *Norms, logics and information systems*, eds. P. McNamara and H. Prakken, 73–90. Amsterdam: IOS Press.
186. Trapp, R.W. 1985. Utility theory and preference logic. *Erkenntnis* 22:301–339.
187. Tversky, A. 1969. Intransitivity of preferences. *Psychological Review* 76:31–48.
188. Uckelman, J. 2009. *More Than the sum of its parts*. PhD thesis, ILLC, University of Amsterdam.
189. Uckelman, J. and U. Endriss. 2008. Preference modeling by weighted goals with max aggregation. In *Proceedings of the 11th international conference on principles of knowledge representation and reasoning (KR-2008)*, eds. G. Brewka and J. Lang, 579–587. Menlo Park, CA: AAAI Press.
190. Velazquez-Quesada, F.R. 2009. Inference and update. *Synthese* 169:283–300.

191. Velazquez-Quesada, F.R. 2011. *Small steps in dynamics of information*. PhD thesis, ILLC, University of Amsterdam.
192. Veltman, F. 1996. Defaults in update semantics. *Journal of Philosophical Logic* 25:221–261.
193. Wedgwood, R. 2003. Choosing rationally and choosing correctly. In *Weakness of will and practical irrationality*, eds. S. Stroud and C. Tappolet, 201–229. Oxford: Oxford University Press.
194. Williamson, T. 1988. First-order logics for comparative similarity. *Notre Dame Journal of Formal Logic* 29(4):457–481.
195. Wooldridge, M. 2000. *Reasoning about rational agents*. Cambridge, MA: The MIT Press.
196. von Wright, G.H. 1951. Deontic logic. *Mind* 60:1–15.
197. von Wright, G.H. 1963. *The logic of preference*. Edinburgh: Edinburgh University Press.
198. von Wright, G.H. 1972. The logic of preference reconsidered. *Theory and Decision* 3:140–169.
199. Yamada, T. 2007. Acts of commanding and changing obligations. In *Proceedings of the 7th workshop on computational logic in multi-agent systems (CLIMA VII)*, eds. K. Inoue, K. Satoh, and F. Toni, 2006. Revised version appeared in LNAI 4371, 1–19. Springer: Heidelberg.
200. Yamada, T. 2008. Logical dynamics of some speech acts that affect obligations and preferences. *Synthese* 165(2):295–315.
201. Yap, A. 2006. Product update and looking backward. Technical Report, PP-2006-39, ILLC, University of Amsterdam.
202. Wang, Y. 2010. *Epistemic modeling and protocol dynamics*. PhD thesis, ILLC, University of Amsterdam and CWI.
203. Zarnic, B. 2003. Imperative change and obligation to do. In *logic, law, morality: Thirteen essays in practical philosophy in Honour of Lennart Aqvist*, eds. K. Segerberg and R. Sliwinski, 79–95. Uppsala philosophical studies 51. Uppsala: Uppsala University.
204. Zhao, Z. 2010. *Combining belief and preference: A complete logic*. Manuscript, Peking University.

Index