LOGICS FOR AGENTS WITH BOUNDED RATIONALITY

Zhisheng Huang

December 18, 2002

LOGICS FOR AGENTS WITH BOUNDED RATIONALITY

Zhisheng Huang

LOGICS FOR AGENTS WITH BOUNDED RATIONALITY

ILLC Dissertation Series 1994-10

For further information about ILLC-publications, please contact

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

Logics for Agents with Bounded Rationality

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam,
op gezag van de Rector Magnificus
Prof.dr P.W.M. de Meijer
in het openbaar te verdedigen in de
Aula der Universiteit
(Oude Lutherse Kerk, ingang Singel 411, hoek Spui)
op donderdag 15 december 1994 te 12.00 uur

door

Zhisheng Huang

geboren te Fujian, China.

Promotor: Dr. Peter van Emde Boas
Faculteit Wiskunde en Informatica
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam

Co-Promotor: Dr. Michael Masuch
Center for Computer Science in Organization and Management (CCSOM)
Faculteit der Politieke en Sociaal-Culturele Wetenschappen
Universiteit van Amsterdam
Oude Turfmarkt 151
1012 GC Amsterdam

Cover design by Zhisheng Huang
ISBN: 90–74795–13–7

Contents

Acknowledgments

First I would like to thank my promotor Peter van Emde Boas. Peter invited me to come to the FWI at University of Amsterdam as a visiting scholar, which started the story of my Ph. D. He encouraged me and provided a pleasant and stimulating atmosphere that gave me a lot of confidence to complete this research. A substantial part of my research consists of joint work. I benefited a lot from Peter's cleverness, kindness and co-operation.

I would like to thank my co-promotor Michael Masuch. Michael invited me to work at CCSOM after my two years' visits to the FWI, which made it possible for this Ph. D. thesis to be completed. During the last four years, Michael gave me his continual support, encouraged me, and taught me how to work out a clear paper. We worked together on action logics for agents with bounded rationality, which constitute the second half of this thesis. I always enjoyed the co-operation with Michael.

The results in the first half of this thesis were obtained during my stay at the FWI. I thank Johan van Benthem and all of the people from the FWI for their support during my visits. In particular, I would like to thank Karen Kwast, Sieger van Denneheuvel, Edith Spaan, Harry Buhrman, Theo Janssen, and Leen Torenvliet who provided a pleasant atmosphere. Special thanks are due to Karen Kwast and Sieger van Denneheuvel for their co-operation with several papers and their kind help in many respects.

This thesis would not have been written the way it is without the stimulating support of László Pólos and Maarten Marx. I co-operated with László on several papers, which appear in the second half of this thesis. Maarten Marx gave me a lot of useful and stimulating suggestions. Thanks are due to Breanndán Ó Nualláin, Gábor Péli, Jeroen Bruggeman, Jaap Kamps, and Babette Greiner for stimulating discussions and conversations on my research. Special thank goes to Jaap Kamps for his kind translation of this thesis' summary into Dutch. I am grateful to Michel Vrinten, Anja Krans, Babette Greiner, Johan Henselmans, Henk Helmantel, and Jeroen van Dellen who provided much kind help.

I am grateful to J.-J. Ch. Meyer for many useful and stimulating comments and

suggestions on my research. Also, thanks to Bernd van Linder, Wiebe van der Hoek, Yao Hua Tan, Gerard Vreeswijk, and Elias Thijsse for their stimulating conversations. I thank Rob Mokken and all of the people from the PSCW for offering a pleasant research and working environment. Special thank goes to Rob Mokken for the 1993 Mokken award for one of my papers, which now constitutes chapters 3 and 4 of this thesis.

Most of all I want to thank my parents for all the support and encouragement they gave me, and who consider their son's earning the title of Ph. D. as one of the happiest events in their lives. I owe so much to my wife Yuanhua and my daughter Lanhong for their love, support, understanding, and for enduring the many years of my absence while we ought to have been together.

Amsterdam Zhisheng Huang
October, 1994.

Chapter 1

Introduction

1.1 Main Work and its Significance

Bounded Rationality has two interpretations: a wide one and a narrow one. In the wide interpretation, *bounded rationality*, or alternatively called, *limited rationality*, refers to the phenomenon agents have limited cognitive resources and capabilities [Doyle 1991]. In the narrow interpretation, bounded rationality refers to the notion raised by H. A. Simon in [Simon 1955]. He considers a general decision procedure for a rational agent who would not know all action alternatives, nor the exact outcome of each, and would lack a complete preference ordering for those outcomes. In this thesis, I propose and study several logics for agents with bounded rationality in both the wide interpretation and the narrow interpretation.

For the wide interpretation of bounded rationality, I focus on the phenomenon of belief dependence in multiple agent environments, where *belief dependence* refers to the phenomenon that some agents rely on someone else about their beliefs, knowledge, or information because of their own limited information or beliefs. In this thesis, the main achievements concerning this perspective are: a general methodology for the study of belief dependence is presented; several logics for belief dependence are proposed; and the soundness, completeness and decidability of those logics are studied.

Moreover, I show that the logic of belief dependence is a useful tool for analyzing the behavior of interacting agents by using the proposed logics to capture a mechanism to guide rational agents to update their information states upon reception of information from other agents. Those approaches have significant application potential in computer science and artificial intelligence. Theoretical analysis of agents who perform updates on their information state upon reception of information from other agents traditionally is based on a number of unrealistic assumptions: there exists a true real state of the world which can be described in terms of a finite set of propositional statements; all agents are correctly informed, even when they have no full information; agents will only provide correct information, and such information will be incorporated correctly upon reception. Even within this idealized scenario

the process of information updating turns out to be non-trivial to describe. The more involved case where agents may start with false information but where only true information is exchanged already leads to the invocation of non-trivial belief revision operators. But real life situations are even less well behaved: agents disagree about the state of the world and will exchange inconsistent information and try to deceive each other. But also the combination of consistent information from two agents may yield unrequired information states, as indicated by the Schoenmakers problem [Schoenmakers 1986] about a judge with two witnesses, whose testimony is correctly combined by the judge into an assertion which they both would have denied.

Any approach in artificial intelligence which attempts to model the behavior of interacting agents which behave in a more human like fashion must therefore separate the process of exchanging information from the incorporation of received information in the agents' belief sets. Moreover, the decision which kind of incorporated operator to invoke under which circumstances should be made effectively computable, in order that a mechanized collection of agents could actually perform the behavior as described by the theory.

Logic of belief dependence provides us with a possible framework for discriminating between the various update operators which an agent may invoke. The knowledge of the agents is extended by information expressing which agent depends on whom with respect to which proposition. This creates the possibility of assigning degrees of credibility to agents generating information, in order that the information provided may more strongly be incorporated the more credible the informant turns out to be.

We also have applied the logic for belief dependence in order to analyse the Schoenmakers problem. The possibilities and impossibilities of strategies for dealing with this problem were investigated. A first attempt tries to characterize those situations where information can be combined without risking the undesirable situation that a derivable proposition contradicts the beliefs of all agents involved: the relevant notion is called *Absolute safety*. It turns out that in that case only trivial solutions exist. It follows that any non-trivial strategy must use additional information about the epistemic states of the agents involved.

Subsequently less restrictive notions of safety were investigated. The more interesting of these notions are those which involve not only propositions about the world but also epistemic information relating the knowledge of the various agents involved. For this purpose the logic of belief dependence is used. The results characterizing the generalized safety notions generalize for this extended logic. A notion of almost safety within this framework which describes the safety of combining information under the hypothesis that agents eventually might have exchanged their information among themselves is presented. For this notion of almost safety an explicit solution to the Judge puzzle is given.

For the narrow interpretation of bounded rationality, I focus on the studies of action logics for agents with H. A. Simon's bounded rationality in order to develop a formal language for social science theories, in particular for theories of organization. These theories are usually expressed in natural language. They lack a formal

foundation that would allow one to check their consistency in a rigorous fashion, or to disambiguate natural language statements. As a consequence, these theories have acquired a reputation for "softness" – a soft way of saying that their logical properties are somewhat dubious. Reformulating them in a formal language with known properties would facilitate the tasks of consistency checking or disambiguation. Also, it would prepare the ground for other tasks, for instance the examination of deductive closure properties.

I focus on action logic, because actions (of individual or collective agents) are key to the understanding of social phenomena. In fact, most social scientists agree that adequate theories of social relations must be action theories first [Blumer 1969, Giddens 1979, Harel 1984, Luhmann 1982, Parsons 1937, Schutz 1967]. Yet actions lead to a change, and change is notoriously hard to grasp in the extensional context of first order languages [Gamut 1990]. This explains our attempt to develop a new logic, rather than taking First Order Logic off the shelf. This new logic is called *ALX* (the x'th action logic).

Action logics are not new in formal AI. There have been a variety of attempts to put actions into a logical framework since McCarthy and Hayes' seminal paper [McCarthy&Hayes 1969], in particular[Cohen&Levesque 1987, Cohen&Levesque 1990, Ginsberg&Smith 1987, Jackson 1989, Rao&Georgeff 1991, Winslett 1988]. If our approach deserves attention, it is because we are knitting together ideas from various strands of thought, notably H.A. Simon's notion of *bounded rationality*, G. H. Wright's approach to *preferences*, Kripke's *possible world semantics* in combination with binary modal operators, Stalnaker's notion of *minimal change*, and more recent ideas from *belief revision* and *update* [Grahne 1991, Hirofumi&Mendelzon 1991]. The resultant logic, called ALX1, contributes a primitive system which can only serve part of our application requirements. However, the soundness and completeness of this primitive system is proved.

Moreover, in order to achieve a formal language which can meet more application requirements, we have to extend the primitive logic into more complicated one, ALX3, or alternatively called *MALX* [1], which includes also the first order logic and belief logics. Moreover ALX3 is a multi-agent ALX logic. Combining so many logics into one makes the resultant logics extremely complicated with the consequence that the decidability is lost. However, we believe that we have reached at least a first formal logic which can serve our primitive application requirements. Therefore, the research work concerning this part is more application-oriented. It represents a trade-off between logical elegance and efficiency of application.

1.2 Organizations of This Thesis

The thesis consists of two parts: Part I "Logics for Belief Dependence" and Part II "Action Logics for Agents with Bounded Rationality". Part I studies the logics for belief dependence, a study of bounded rationality in the wide interpretation. Part

[1] There exists an ALX2 logic in our research, which is a simplified and intermediary version of the first order ALX logic. For the details, see [Huang, Masuch&Pólos 1993].

II is a study of the action logics for agents with bounded rationality in the narrow sense.

In Part I, the chapter "Bounded Rationality and Belief Dependence" reviews the notions of rationality and limited rationality, examines the variants of limited rationality, and discusses the significance of their application. We present evidence for the conclusion that belief dependence is one of main phenomena of limited rationality. Also, the phenomenon of belief dependence is systematically examined. We conclude that compartment belief plays an important role in the study of belief dependence. Our methodology for belief dependence logics is based on this observation. The syntax for these logics is considered and a general scenario for the formalization of belief dependence is presented.

In the chapter "Formalizing Belief Dependence", several logics for belief dependence are proposed. We can see that the proposed logics are sufficiently expressive to formalize the communication of information among multiple agents with limited information.

The chapter "Semantics Models" introduces the semantic models for the logics. I argue that general epistemic logics and doxastic logics are not appropriate tools for logic of belief dependence. Subsequently, I propose several semantic models, compare those models, and discuss when and under which situations these semantic models are suitable for application.

In the chapter "Belief Dependence, Revision, and Persistence", I study the belief dynamics in the framework of belief dependence. Using the belief dependence logics, I offer a mechanism to calculate how an agent can make a choice between various belief update alternatives like belief revision, expansion, contraction, and persistence.

In the chapter "Information Acquisition from a Multi-agent Environment", I study a problem originally proposed by W. Schoenmakers, which represents a typical example in the study of information acquisition from multi informants. A general approach for formalizing the problem of information acquisition from multiple sources is presented. Several notions which are motivated from Schoenmakers problem, such as absolute safety, safety, and strong safety are formally defined. Moreover, using the logic of belief dependence, a general notion of almost safety is defined, which is shown to be a reasonable and acceptable strategy for the Schoenmakers problem.

In the chapter "Conclusions and Further Work", further work on the belief dependence logic is discussed, and concluding remarks for this part are stated.

Part II starts with the chapter "Introduction" where we present general ideas about action logics for agents with bounded rationality as a formal language for social science theory, (specially, the theory of organizations). It is argued that action logic plays an important part in reasoning about organizations. Moreover, we explain why H. A. Simon's notion of bounded rationality, G. H. v. Wright's approach to preference, S. Kripke's possible world semantics, Stalnaker's notion of minimal change, and update semantics have been combined into our system for action logic.

In the chapter "Preference Logics", we examine the notion of preference, and distinguish four kinds of preference relations, called *actual preference, ceteris paribus preference, conditional preference*, and *absolute preference*. Moreover, I provide both

syntactic and semantic characterizations for each of them. Furthermore, the correspondings preference logics are considered.

In the chapter "ALX1: A Propositional ALX Logic", I propose a primitive system combining preference logic, update logic, and propositional dynamic logic. The soundness and completeness is proved. Furthermore, minimal change actions are studied.

The chapter "ALX3: A Multi-agent ALX Logic" deals with a multimodal predicate version of ALX logic. ALX3 extends ALX1 by the first order logic, more action combinations, belief operation. In the chapter, we will indicate some features applying multi-agents action logic, discuss some of plausible applications of the ALX logic towards a formal theory of social agents, and show that ALX3 logic is indeed sufficiently (in some sense) expressive to serve our application purposes. A final concluding remarks completes this thesis.

Part I

Logics for Belief Dependence

Chapter 2

Bounded Rationality and Belief Dependence

2.1 Bounded Rationality: the wide interpretation

The notion of *rationality* in decision theory involves choosing the right action by an agent with given her preferences and the outcomes of these actions. Usually, a choice is said to be rational for an agent if it is of maximal expected utility with respect to the agent's beliefs and preferences. The notion of rationality in logic and artificial intelligence concerns correct capability in making reasoning.

Idealed with the capability of obtaining by reason the correct choice, traditional decision theory and logical theory have been based on the assumption that agents have an *idealized rationality*. In decision theories, the idealized rationality assumes that agent possesses a full knowledge about preferences and outcomes. In logic, computer science, and artificial intelligence, the idealized rationality implies that agents are logical omniscient, have complete and consistent knowledge, and have unlimited cognitive resources and capabilities. A large amount of existing work in logic, computer science, and artificial intelligence has, implicitly or explicitly, assumed that intelligent agents possess this idealized form of rationality.

Traditional theories based on the idealized rationality assumption face serious difficulties in the applications, since in real life, both human beings and artificial agents (computers, knowledge bases, robots, and processes), are non-idealized agents. Therefore, researchers in logic, artificial intelligence, computer science, and decision theories, are seeking for approaches where the idealized rationality assumptions are somewhat weakened, loosened, or even completely removed. This results in a proliferation of new theories and techniques oriented towards bounded rationality.

As it is mentioned above, the notion of *bounded rationality*, or alternatively called *limited rationality*, has a wide interpretation and a narrow interpretation. In this part, we let the term "bounded rationality" refer to the phenomenon that agents have limited cognitive resources and abilities. We distinguish the following aspects of the bounded rationality in the wide interpretation.

1) **Incomplete Information**
The rational agents have incomplete information, beliefs, preferences, and knowledge. Partial logics [Thijsse 1992] and situation semantics [Barwise&Perry 1983] are formal tools to study this phenomena of incomplete information.

2) **Inconsistent Information**
The rational agents have contradictory beliefs, incompatible preferences, and inconsistent knowledge. However, the inconsistency of information does not necessarily imply the agents' state of mind reduces to absurdity. It might be the case that agents distribute their inconsistent beliefs into different mind frames. Jon Doyle's society of mind theory offers a framework to deal with the problem [Doyle 1983].

3) **Limited Resources**
The rational agents have limited time to solve problems. They have limited memory to remember information. They lack certain cognitive capabilities to cope with some difficulties. Levesque's logic for implicit belief and explicit belief, Fagin and Halpern's awareness logic, and their relevant approaches aimed at solving the problem of logical omniscience all fall into the category of approaches which involve both incomplete information and limited resources.

The logic of belief dependence which is studied in this thesis is also oriented towards bounded rationality. Compared to the existing other bounded rationality approaches, the logic of belief dependence has some novelties: the proposed logic focuses on the phenomena of belief dependence among multiple agents, and the logic offers a more powerfully formal tool in reasoning about knowledge and belief which originates from other agents. One of the most important topics in artificial intelligence and computer science, when studying bounded rationality, is the problem of reasoning about knowledge in a multi-agent environment. Reasoning about knowledge in such environments has already found many applications e.g., distributed knowledge-bases, communication, and cooperation for multi-agent planning [Bond&Gasser 1988, Halpern&Fagin 1989, Cohen&Levesque 1987, Levesque 1984] [Werner 1988]. However, in these existing approaches, little attention has been paid to the problem of belief dependence in multi-agent environments, where agents may rely on others for their beliefs and knowledge because their own information is limited. In multiple agent environments, it is frequently beneficial to have agents communicate their knowledge to others, because individual resources are limited, so division of activity may help. Although there have been attempts to study the problem of the communication of belief and knowledge among multiple agents [Fagin&Vardi 1986, Halpern&Fagin 1989, Werner 1988], the existing formalisms generally focus on the problem of communication, and consequently some important features of belief dependence, such as suspicion and indirect dependence, are rarely formalized. The logic of belief dependence offers a formal tool to formalize these aspects of belief dependence, and may provide a foundation for understanding the phenomena of the belief communication among multiple agents. Moreover, the pro-

posed formalism may find also applications in fields such as knowledge acquisition, machine learning, human-computer interaction, distributed artificial intelligence and distributed network systems.

In this chapter, I will first examine the problem of belief dependence in details, and then discuss some crucial notions concerning belief dependence. Next, we consider the syntax for the logics of belief dependence.

2.2 Belief Dependence

2.2.1 Compartmentalized Information and Incorporated Information

As mentioned above, both human beings and artificial agents are of bounded rationality. In other words, rational agents have only limited resources; they have only limited time to solve problems; they have limited memory to remember information; they lack certain capabilities to cope with some difficulties, and they have insufficient knowledge to fulfil some tasks. However, in order to show their rationality, to make themselves more flexible, to prove their intelligence, and even, to keep alive, these rational agents must seek for help from other agents in a multi-agent environment. They may convey problems to other rational agents to solve. They may ask for support from others in order to extend their capabilities. Frequently, they seek just some information from other rational agents.

When a rational agent, say i, seeks some information from some other rational agent, say j, we can say that the agent i has some belief dependence on the other agent j. We call the phenomena in which some agent depends on some other agents for their knowledge and beliefs *belief dependence*. When a rational agent receives some information from others, they may commit to several different strategies to handle the received information. In epistemology, there have been many studies on this issue. Among these I mention *the foundation theory* and *the coherence theory*.

According to the foundation theory, one needs to keep track of the justification for one's beliefs. One accepts only information with such a justification as her own knowledge and beliefs. According to the coherence theory, one needs not consider the origin of one's beliefs. One assimilates the new beliefs which are coherent with one's original beliefs. The foundation theory is a strong version of the coherence theory, because if a belief is justified, the belief must be coherent. In other words, incoherent belief cannot be viewed as a justified belief.

In real life, rational agents are neither pure foundationists nor pure coherencists. Frequently, they behave as controled by a mixture of both theories. In fact, there might exist a third option leading to a theory, called the *compartment theory*. Consider the situation where some agent i receives the information ϕ from others, and ϕ is coherent with her original beliefs, but not justified. Moreover, the agent i does not intend to commit herself as a pure foundationist or as a pure coherencists. Therefore, she cannot accept the information ϕ, although ϕ is coherent, but ϕ is not justified. Moreover, she also cannot refuse ϕ, although ϕ is not justified, since ϕ is coherent.

Under this situation, a rational strategy to cope with the problem is to just keep ϕ under a compartmentalized status, namely, ϕ is kept as a sub-belief, which is neither a completely accepted one, nor is a completely refused one. This sub-belief is called a *compartmentalized belief*.

Therefore, we see that the compartment theory presents an intermediate strategy between the foundation theory and coherence theory. The compartment theory is also a reasonable and intuitive strategy to formalize the dynamics of rational agents' beliefs.

In their study of incorporating new information into existing world knowledge of human beings, cognitive psychologists also make a distinction between compartmentalized information and incorporated information. As Potts et al. point out in [Potts et al., 1989]:

> ...it is unlikely that subjects in most psychology experiments incorporate the new information they learn into their existing body of world knowledge. Though they certainly use their existing world knowledge to help comprehend the new material, the resulting amalgam of new information, and the existing world knowledge used to understand it, is isolated as a unit unto itself: it is *compartmentalized*.

I also believe that an appropriate procedure to assimilate others' knowledge and beliefs should consists of the following two phases: one producing compartmentalized information and another one leading to incorporated information. Formally, *compartmentalized information* consists of those fragments of information which are accepted and remembered as isolated beliefs (and which are somewhat different from those beliefs which are completely believed), whereas *incorporated information* consists of those beliefs which are completely believed by the agents.

Compartmentalized belief may be understood in the following different manners:

1) **Society of Minds**
The notion of *society of minds* [Doyle 1983] is that each agent possesses its own separated cluster of beliefs, which may be contradictory. Each cluster of beliefs is connected with some mind frame. However, if accepted information is simply scattered across in different mind frames, it is hard to say that an agent can assimilate others' beliefs efficiently and can enlarge her belief set.

2) **Probability-based Beliefs**
One might consider the use of subjective probability as another sort of compartmentalization. However, it is rather strange to consider the belief that it is raining with 65% likelyhood to be separated from the complementary belief that it is not raining with 35% likehood. Compartimentalization rather refers to the possible alternative assignments of subjective probabilities which may be the result of having heard the weather forecast on the radio. The possibility assignment to proposition should be treated into a united object, rather than a partitioning of the frame of mind.

3) **Source Indexing**

By *source indexing* we mean that the received information is indexed with the name of the informant. The compartmentalized belief can be understood as one which has source indexing. In the following, I will argue that the source indexing plays an important part in the formalization of belief dependence. Therefore, in the logic of belief dependence, compartmentalized beliefs are the sub-belief which are indexed by informants.

2.2.2 The Roles of Source Indexing of Information

Rational agents (human beings, computers, ect.) often receive information from outside. When a rational agent receives some information from outside, she can think that she gets the information from an *agent*, by viewing any entity which can bring about her receiving information as an agent. Moreover, when a rational agent receives some information from other agents, normally, she receives not only the information, but also knows the fact from whom she gets such information, that is, some additional information about the informant.

Sometimes the agent has no exact knowledge about the informant, when she receives some information. However, she may have some beliefs about the informant. Moreover, when a rational agent receives the additional information, the additional may not only include the information about the informant, but also the time when the information is received, and the location where the information is received, and the attitude of the informant. However, here I just focus on the problem in which additional information is about informants.

The phenomenon that rational agents always receive some additional information about the informant when receiving information, is ubiquitous. When you hear the news "The Soviet Union does not exist any more", you know that you receive the news from BBC, or CNN, from your wife, even your kids, or from a stranger who is talking with someone else in the public toilet. When a knowledge base system receives an input information "$\pi = 3.1415$", the knowledge base system may only know that the information comes from a terminal, say, named TA, which is connected with the system. If we want to develop a more intelligent man-machine interaction system for knowledge bases, we must rely on such a logic of belief dependence, by which knowledge systems can distinguish different users who try to control the knowledge bases by inputing some information; and knowledge base systems can actually be programmed to obey only some authorized agents and refuse information from agents with "evil" intentions.

In [Gabbay 1992], Dov Gabbay offers several interesting examples which concern reasoning about informants.

2.2.1. Example. (Jethrow's Career) *The figure "Jethrow's Career S" is a database S with source indexing about Jethrow's performance. It indicates the source supporting the truth of the predicate. The following database lists candidates for directorship of a new Max-Planck Institute in Germany. The database is labeled mp. It contains data about candidates labeled by their source, and some non-numeric evaluation.*

2.2.2. Example. (Dov's Buying a House) *Let $B(a)$ be a literal meaning "It is a*

	$Student:$	$goodteacher(J)$,
	$survey$	
	$Letters:$	$goodresearch(J)$,
$S:$	$Students:$	$fatherlyfigure(J)$

Figure 2.1: Jethrow's Career S

	$(fairlystrong, S):$	$candidate(J)$,
$mp:$	$(preferred, T):$	$candidate(H)$.

Figure 2.2: Jethrow's Career mp

sound investment for Dov to buy the house a for the price quoted". The words in the figure indicate source of information. The accountant and lawyer recommend that Dov buys the house. So does Dov and so does his wife. Dov likes it. The accountant thinks that Dov has the money and that it is a good move. The lawyer checked with his assistant the legal aspects and interviewed his informer at City hall. The area development plan looks good. So everybody agrees that $B(a)$ is true except for the mother-in-law, who for her own (non-logical) reasons says no. The figure 2.3 represents a belief dependence database describing the above state of affairs. A further mechanism (logical or decision theoretical) to draw a conclusion from this database is needed.

2.3 Logics of Knowledge and Belief and Logical Omniscience

2.3.1 General Logic of Knowledge and Beliefs

In philosophy logic of knowledge is called *epistemic logics*, whereas logic of beliefs is called *doxastic logic*. Possible worlds semantics was first proposed by [Hintikka 1962] for models of the logic of knowledge and belief. The intuitive idea beyond possible worlds semantics is that besides the actual world, there are a number of other possible worlds, or states. Some of those possible worlds may be indistinguishable for an agent from the actual world. An agent is said to know a fact φ if φ is true in all the worlds she thinks possible.

In this section we briefly review the possible worlds semantics for knowledge and belief. Suppose we consider a logic system involving n agents, where $\mathbf{An} = \{i_1, ..., i_n\}$ denotes the set of agents, and where we have a set Φ_0 of primitive propositions about which we wish to reason. In order to formalize the reasoning about knowledge and belief, we use a modal propositional logic, which consists of the standard connectives such as $\wedge, \vee$,and$\neg$, and some modal operators $L_i, L_j, ...,$. A formula such as $L_i\varphi$ is to be read as "agent i knows that φ" if we interpret the operator as a knowledge

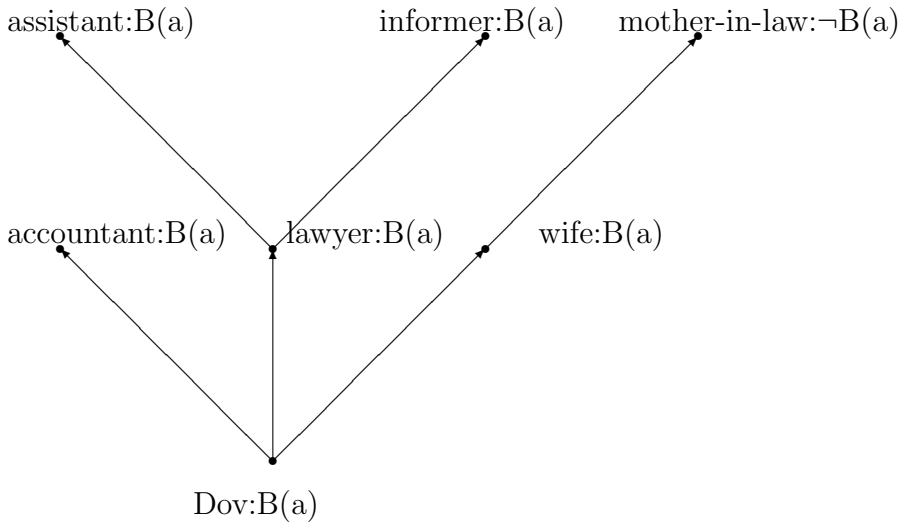


Figure 2.3: Dov's Buying a House

operator, or "agent i believes φ" if we interpret the modal operator L_i as a belief operator.

We give semantics to these formulas by means of Kripke structures, which formalize the intuitions behind possible worlds. A Kripke structure for knowledge for n agents is a tuple $\langle W, \mathcal{L}, V\rangle$, where W is a set of possible worlds, V is a truth assignment which assigns to each primitive proposition in Φ_0 a subset of possible worlds, and $\mathcal{L} : \mathbf{An} \rightarrow \mathcal{P}(W \times W)$ specifies n binary accessibility relations on W, where $\mathcal{P}$ means the power set. For a knowledge system, the binary relations are equivalence relations. For a belief system, the relations are serial, transitive, and Euclidean. A relation R on a set S is *serial* if for each $s \in S$ there is some $t \in S$ such that $\langle s, t\rangle \in R$; R is *transitive* if $\langle s, u\rangle \in R$ whenever $\langle s, t\rangle \in R$ and $\langle t, u\rangle \in R$; R is *Euclidean* if $\langle t, u\rangle \in R$ whenever $\langle s, t\rangle \in R$ and $\langle s, u\rangle \in R$.

We now assign truth values to formulas at a possible world in a structure. We write $M, w \models \varphi$ if the formula φ is true at possible world w in structure M.

$M, w \models p$, where p is a primitive proposition, iff $w \in V(p)$
$M, w \models \neg\varphi$ iff $M, w \not\models \varphi$,
$M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$,
$M, w \models L_i\varphi$ iff $M, t \models \varphi$ for all t such that $\langle w, t\rangle \in \mathcal{L}(i)$.

We say a formula φ is *valid in structure* M if $M, w \models \varphi$ for all possible worlds w in M; φ is *satisfiable in* M if $M, w \models \varphi$ for some possible worlds in M. We say φ is *valid* if it is valid in all structures; φ is *satisfiable* if it is satisfiable in some Kripke structure.

The logic of belief above is characterized by the following axiom system, called weak **S5** or **KD45**.

(BA) All instances of propositional tautologies.
(KL) $L_i\varphi \wedge L_i(\varphi \rightarrow \psi) \rightarrow L_i\psi$.
(DL) $\neg L_i \bot$.
(4L) $L_i\varphi \rightarrow L_iL_i\varphi$.
(5L) $\neg L_i\varphi \rightarrow L_i\neg L_i\varphi$.
(MP) $\vdash \varphi, \vdash \varphi \rightarrow \psi \Rightarrow \vdash \psi$.
(NECL) $\vdash \varphi \Rightarrow \vdash L_i\varphi$.

(BA) and (MP) hold from propositional logic. (KL) means that an agent's belief is closed under implication, (DL) says that an agent never believe things that are false. This axiom is generally taken to distinguish belief from knowledge. For a knowledge system, (DL) is replaced by a stronger axiom (TL) $L_i\varphi \rightarrow \varphi$, which says that an agent only knows things that are true. (4L) is the axiom of positive introspection, which mean that each agent believes what she believe. (5L) is the axiom of negative introspection, which says that each agent knows what she does not believe. (NECL) is a generalization rule for the operator L_i, which says that the agent believes all of tautologies. It is responsible for one of the problems of logical omniscience which are examined in the details in the next subsection.

2.3.2 The Problem of Logical Omniscience

Possible world semantics for knowledge and belief does not seem to be an appropriate theory for modelling human reasoning, because it suffers from *the problem of logical omniscience*. An agent is *logical omniscient* if, whenever she believes all of the formulas in a set Ψ, and Ψ logically implies the formula φ, then the agent also believes φ. It is well known that humans, or even computers, are not such perfect reasoners, because they are generally of bounded rationality. In other words, these epistemic logics rather capture logically possible knowledge and beliefs instead of the agents' actual knowledge and beliefs.

To solve the problem of logical omniscience, one of the approaches is to focus on the invalidation of some logical closure by a logical strategy. Formally, we can formalize the closure properties as follows. Let Ψ_K be a set of formulas for an epistemic or doxastic modal operator K. For a semantics model M, the modal operator K is said to be:

(C1) *closed under implication,*
if $\varphi \in \Psi_K$, and if $\varphi \rightarrow \psi \in \Psi_K$, then $\psi \in \Psi_K$.
(C2) *closed under conjunction,*
if $\varphi \in \Psi_K$, and $\psi \in \Psi_K$, then $\varphi \wedge \psi \in \Psi_k$.
(C3) *decomposable under conjunction*
if $\varphi \wedge \psi \in \Psi_K$, then $\varphi \in \Psi_K$, and $\psi \in \Psi_K$
(C4) *closed under axioms of logical theory T,*
if φ is an axiom of some logical theory T, then $\varphi \in \Psi_K$.
(C5) *closed under valid formula,*
if φ is a tautology, then $\varphi \in \Psi_K$.

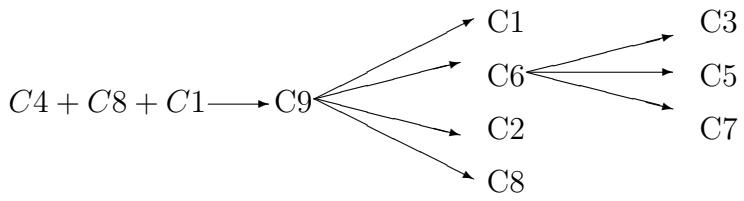


Figure 2.4: Dependencies between Closure Conditions

(C6) *closed under valid implication,*
if $\varphi \in \Psi_K$, and if $\varphi \to \psi$ is valid, then $\psi \in \Psi_K$.
(C7) *closed under logical equivalence,*
if $\varphi \in \Psi_K$, and φ is logically equivalent to ψ, then $\psi \in \Psi_K$.
(C8) *closed under substitution,*
if $\varphi \in \Psi_K$, then $\varphi\theta \in \Psi_K$ for any substitution θ.
(C9) *logical omniscience,*
if Ψ_K logically implies φ, then $\varphi \in \Psi_K$.

There exist at least the following dependencies among those closure conditions above.

(a) $C1 + C5 \to C6$.
(b) $C4 + C8 + C1 \to C9$.
(c) $C5 \to C4$.
(d) $C6 \to C3 + C5 + C7$.
(e) $C9 \to C1 + C6 + C2 + C8$.

Generally, the fewer closures are implied, the more acceptable the condition is. From the dependencies between closure conditions above, we know that (C1), (C6) and (C2) play an important part in the problem of logical omniscience. Consequently the existing approaches primarily focus on those three closure properties.

2.3.1. Claim. *In general epistemic logics, beliefs are closed under implication, valid implication and conjunction.*

There are some proposals which introduces the notion of *non-classical worlds* in the semantics to solve the problem of logical omniscience. *Non-classical worlds* are worlds in which not all valid formulas need be true. Moreover, in non-classical worlds some inconsistent formulas may be true, hence they are called *impossible worlds* or *nonstandard worlds*.

In [Levesque 1984], Levesque first proposed the notions of implicit and explicit belief. Formally, Levesque uses two modal operators B and L to stand for explicit belief and implicit belief respectively. A *structure for implicit and explicit belief* is a tuple $M = \langle S, \mathcal{B}, T, F\rangle$, where S is a set of situations, $\mathcal{B}$ is a subset of S, and $T, F : \Phi_0 \to \mathcal{P}(S)$. Intuitively, $T(p)$ consists of all situations that support the truth of p, whereas $F(p)$ consists of all situations that support the falsity of p. Obviously, in a situation, a proposition may be true, false, both, or neither. Situations which

supports neither the truth nor falsity of some primitive proposition are called *partial situations*. An *incoherent situation* is the situation which supports both the truth and falsity of some primitive propositions.

A *complete situation*, or a possible world, is one that supports either the truth or falsity for every primitive proposition and is not incoherent. A complete situation is *compatible* with a situation s' if $s' \in \mathrm{T(p)}$ implies $s \in \mathrm{T(p)}$, and $s' \in \mathrm{F(p)}$ implies $s \in \mathrm{F(p)}$, for each primitive proposition p. $\mathcal{B}^*$ stands for the set of all complete situations in S compatible with some situations in $\mathcal{B}$.

Now, we can define the *support relations* $\models_T$ *and* $\models_F$ between situations and formulas as follows:

$M, s \models_T p$,	iff $s \in T(p)$, where p primitive,
$M, s \models_F p$,	iff $s \in F(p)$, where p primitive;
$M, s \models_T \sim \varphi$	iff $M, s \models_F \varphi$,
$M, s \models_F \sim \varphi$	iff $M, s \models_T \varphi$;
$M, s \models_T \varphi_1 \wedge \varphi_2$	iff $M, s \models_T \varphi_1$ and $M, s \models_T \varphi_2$,
$M, s \models_F \varphi_1 \wedge \varphi_2$	iff $M, s \models_F \varphi_1$ or $M, s \models_F \varphi_2$;
$M, s \models_T B\varphi$	iff $M, t \models_T \varphi$ for all $t \in \mathcal{B}$,
$M, s \models_F B\varphi$	iff $M, s \not\models_T B\varphi$;
$M, s \models_T L\varphi$	iff $M, t \models_T \varphi$ for all $t \in \mathcal{B}^*$,
$M, s \models_F L\varphi$	iff $M, s \not\models_T L\varphi$.

From the definitions above, it is ease to see that explicit belief implies implicit, namely, the following axiom holds:

$$\models (B\varphi \rightarrow L\varphi).$$

Moreover, although implicit belief is closed under implication and valid implication, explicit belief does not suffer from the problem of logical omniscience.

2.3.2. Claim. *In Levesque's explicit and implicit beliefs logic, explicit beliefs are closed and decomposable under conjunction, but they are neither closed under implication, nor closed under valid implication.*

As Levesque points out, the following axiom is valid in Levesque's semantics:

$$B\varphi \wedge B(\varphi \rightarrow \psi) \rightarrow B(\psi \vee (\varphi \wedge \neg\varphi)).$$

This means that either the agent's beliefs are closed under implication, or else some situation he believes possible is incoherent. Imagining that an agent could consider an incoherent situation possible is generally against our intuitions. Also, Levesque' explicit and implicit logic suffers from a critical representation problem since the language is restricted to formulas where no B or L appears within the scope of another.

Another approach to solve the problem, which is often called a syntactic approach, describes an agent's original actual beliefs by a set of formulas, called the *base beliefs set*, and obtains the logical consequences of the base beliefs set by using some logically incomplete deduction rules.

In [Konolige 1983], Konolige presents a *Deductive Belief System,* in which an agent's beliefs are described as a set of sentences in some formal language, together with a deductive process for deriving consequence of those beliefs. In Konolige's deductive belief system, the general model of deduction is a block tableau sequent system. A block tableau system τ consists of a set of axioms and deduction rules. Konolige's Deductive beliefs model can account for the effect of resource limitations on deriving consequences of the base set. As a consequence, an agent need not believe all the logical consequences of his beliefs.

However, syntactic approaches are generally difficult for analyzing the properties of knowledge and belief, since knowledge and beliefs are simply represented by an arbitrary set of formulas. For artificial agents such as robots, computers, or knowledge-bases, deduction models of beliefs may be reasonable. However, for rational agents such as humans, even intelligent artificial agents, beliefs obtained by deduction models still are viewed as logically possible beliefs instead of actual beliefs since in rational reasoning there seems to be no simple logical deduction closure for their actual beliefs at all.

In [Fagin&Halpern 1988], Fagin and Halpern point out that an agent's lack of knowledge of valid formulas is not due to incoherent situations, but is rather due to the lack of "awareness" on the part of the agent of some primitive propositions, and similar reasons hold for the lack of closure under valid implication.

In order to solve the problem of awareness, Fagin and Halpern offer a solution where one can decide on a metalevel what formulas an agent is supposed to be aware of. They provide a logic of general awareness, which can be viewed as an approach which combines the syntactic approaches and nonclassical worlds approaches.

In Fagin and Halpern's general awareness logic, in addition to the modal operator B_i and L_i of Levesque's Logic, they also use a modal operator A_i for each agent i. They give the formula $A_i\varphi$ a number of interpretations: " i is aware of φ," "i is able to figure out the truth of φ," or even in the cases of knowledge bases," agent i is able to compute the truth of φ within time T."

Supposed we have a set $\mathbf{A_n}$ of agents and a set Φ_0 of primitive propositions. Let $\mathbf{L_M}$ be the formula set which is generated recursively from the primitive propositions, Boolean connectives and model operators as usual. A *Kripke structure for general awareness*[1] is a tuple:

$$M = \langle W, \mathcal{L}, \mathcal{A}, V \rangle$$

where W is set of possible worlds, or called *states*,
V is a truth assignment for each primitive proposition $p \in \Phi_0$,
$\mathcal{L} : \mathbf{A_n} \rightarrow \mathcal{P}(W \times W)$, which consists of n serial, transitive, Euclidean relations on W,
$\mathcal{A} : \mathbf{A_n} \times W \rightarrow \mathcal{P}(\mathbf{L_M})$, is a function which assigns to each agent and each world a formula set.

[1]Here the notations are different from Fagin and Halpern's original ones because we would like to preserve notational consistency.

Truth conditions are:

$$
\begin{array}{lll}
M, w \models p & \text{iff} & w \in V(p), \text{ where } p \text{ primitive,} \\
M, w \models \neg\varphi & \text{iff} & M, w \not\models \varphi, \\
M, w \models \varphi_1 \wedge \varphi_2 & \text{iff} & M, w \models \varphi_1 \text{ and } M, w \models \varphi_2, \\
M, w \models A_i\varphi & \text{iff} & \varphi \in \mathcal{A}(i, w), \\
M, w \models L_i\varphi & \text{iff} & M, w' \models \varphi \text{ for all } w' \text{ such that } \langle w, w' \rangle \in \mathcal{L}(i), \\
M, w \models B_i\varphi & \text{iff} & \varphi \in \mathcal{A}(i, w) \text{ and } M, w' \models \varphi \text{ for all } w' \text{ such that } \langle w, w' \rangle \in \mathcal{L}(i).
\end{array}
$$

2.3.3. Claim. *In Fagin and Halpern's general awareness logic, explicit beliefs are not closed under conjunction, neither decomposable under conjunction, nor closed under implication, nor closed under valid implication.*

The general awareness logic has the property that agents are not logically omniscient, and the logic is more suitable than traditional logics for modelling beliefs of humans (or machines) with limited reasoning capabilities.

In general, the awareness in the logic can be viewed as a complex psychological function which integrates other relevant psychological and computational factors such as attention, prejudices, reasoning capabilities, etc.

In order to capture a more intuitive understanding of the notion of awareness, it is necessary to make a detailed analysis on awareness. There seem to exist many kinds of interpretations of the notion of awareness.

Awareness by perception

A simple psychological interpretation is *awareness by perception*, which says that to be aware of something is to perceive something. Awareness of a compound is generally[2] built up from the awareness of its parts, namely, the perception of its parts. A suitable semantics approach to formalize the awareness by perception seems to be situation semantics which is proposed by [Barwise&Perry 1983].

Awareness by computation

Another interpretation of awareness is awareness by computation, which means that to be aware of something is to be able to figure out the truth of that by some special deduction rules or other approaches. In other words, non-awareness of something can be interpreted as failure of computation of the truth. That may be because the agent's resources are limited or for some other reason. From the computational point of view, as suggested in [Konolige 1986], there are two possible approaches that would fit into the awareness framework:

1. *Awareness as filter*
   Agents compute all logical consequences of their beliefs, throwing away those not in the awareness set, perhaps because limitation of memory, perhaps because of agents' prejudices.
2. *Awareness as derivator*
   Agents use a complete logical deduction system to compute consequences of

[2] As was pointed out to me by John-Jules Meyer, there are examples where one perceives the whole but not the components.

beliefs, but do not pursue those lines of reasoning which require deriving sentences not in the awareness set.

Indirect awareness

An agent may be not directly aware of a formula φ. But he may be aware of the agents who are aware of φ. That case is called *indirect awareness*. In general awareness logic an agent cannot have any explicit belief about some formula φ if he is not aware of φ. However, we argue that unawareness does not necessarily result in the failure of capturing explicit beliefs. For instance, suppose you read a sentence says '*The rabbit is an oryctolagus cuniculus*' in a zoological book. Although you may not be aware of 'oryctolagus cuniculus', you may believe that the rabbit is an oryctolagus cuniculus is true since you generally believe what the author says. Therefore, indirect awareness can be viewed as an intuitive extension of the notion of awareness. Therefore, the logic of belief dependence offers an alternative to formalize the indirect awareness.

System awareness

In reasoning about multi-agents' knowledge and belief, a reasoner may be one of those agents whose knowledge and belief are formalized and reasoned with. However, specially, the reasoner may not be one of those agents but only an observer, or called *super-agent*. In general awareness logics, we can only formalize on agents' general awareness. It is frequently beneficial to be able to formalize super-agent's awareness itself. We call it *system awareness*. In fact, system awareness set specifies the opinions of a system reasoner about its reasoning capabilities. The notion of system awareness has a close relationship with the notion of "unknown", which is generally introduced in some systems of knowledge and belief, especially, in knowledge bases and the database system with null values [Kwast 1992]. Intuitively, nonawareness means "unknown".

Moreover, in a multi-agents environment, even though a super-agent is one of those agents whose beliefs and knowledge are reasoned with, we should draw a distinction between awareness of a general agent and the awareness of a super agent.

2.4 Syntactic Considerations for Logics of Belief Dependence

We use general epistemic and doxastic operators to represent agents' knowledge and beliefs. For the sake of convenience, just as in general epistemic logics, we use $L_i\varphi$ to represent that agent i knows or believes the formula φ. As is well known, L is interpreted as an epistemic operator, if the logic system is an **S5** system, whereas L is a doxastic operator if the system is a weak **S5** system [Hintikka 1962].

In order to formalize compartmentalized information and source indexing, we introduce a *compartment modal operator* $L_{i,j}$. Intuitively, we can give $L_{i,j}\varphi$ the inter-

pretation: "agent i believes φ due to agent j". From the point of view of Minsky's society of minds, $L_{i,j}\varphi$ can also be intuitively interpreted as "agent i believes φ on the mind frame indexed j". Alternatively we call $L_{i,j}\varphi$ agent i's *sub-belief*, and $L_{i,j}$ is called *sub-belief operator*.

Both sub-beliefs and general beliefs have close relationships with the truth and falsity of beliefs. Sometimes we need a neutral[3] modal operator $D_{i,j}$ for belief dependence logics. $D_{i,j}$ is called the *dependency operator*, or alternatively the *rely-on operator*. Intuitively, we can give $D_{i,j}\varphi$ a number of interpretations: "agent i relies on agent j for the formula φ", "agent i depends on agent j about believing φ", "agent j is the credible advisor of agent i about φ", even especially in distributed process networks, "processor i can obtain the knowledge about φ from processor j". There are two kinds of interpretations for the dependency operator $D_{i,j}$. One is *explicit dependence*, which says that belief dependence is explicitly known by believers. In other words, that means the axiom $D_{i,j}\varphi \rightarrow L_i D_{i,j}\varphi$ holds. The other one is *implicit dependence*, in which believers do not necessarily know their dependencies.

It should be noted that $L_{i,i}\varphi$ is not necessarily equal to $L_i\varphi$. We have argued that the notion of belief dependence can be viewed as an intuitive extension of the notion of awareness, since one can define $A_i\varphi \stackrel{\text{def}}{\Longleftrightarrow} \exists j D_{i,j}\varphi$. This means that agent i is aware of φ if and only if agent i believes φ or agent i could decide whether or not φ is true by asking somebody else. Alternatively one can define $A_i\varphi \stackrel{\text{def}}{\Longleftrightarrow} D_{i,i}\varphi$ suggesting that $L_{i,i}\varphi$ is not necessarily equal to $L_i\varphi$. From the point of view of explicit beliefs and implicit beliefs, $L_i\varphi$ can be interpreted as implicit belief, whereas $L_{i,i}\varphi$ can be interpreted as explicit belief if one defines $L_{i,i}\varphi \stackrel{\text{def}}{\Longleftrightarrow} D_{i,i}\varphi \wedge L_i\varphi$.

Supposed we have a set $\mathbf{A_n}$ of n agents, and a set Φ_0 of primitive propositions, the language $\mathbf{L}$ for belief dependence logics is the minimal set of formulas closed under the following syntactic rules:

(i) $p \in \Phi_0 \Rightarrow p \in \mathbf{L}$,
(ii) $\varphi \in \mathbf{L}, \psi \in \mathbf{L} \Rightarrow \varphi \wedge \psi \in \mathbf{L}$,
(iii) $\varphi \in \mathbf{L} \Rightarrow \neg\varphi \in \mathbf{L}$,
(iv) $\varphi \in \mathbf{L}, i \in \mathbf{A_n} \Rightarrow L_i\varphi \in \mathbf{L}$,
(v) $\varphi \in \mathbf{L}, i, j \in \mathbf{A_n} \Rightarrow L_{i,j}\varphi \in \mathbf{L}$
(vi) $\varphi \in \mathbf{L}, i, j \in \mathbf{A_n} \Rightarrow D_{i,j}\varphi \in \mathbf{L}$

Logical connectives such as $\rightarrow$ and $\vee$ are defined in terms of $\neg$ and $\wedge$ as usual, $\top$ is defined as $\varphi \vee \neg\varphi$ for some formula φ, and $\bot$ is an abbreviation of $\neg\top$.

In some special belief dependence logics, some belief dependency operators are defined in terms of others. For example, the sub-belief operator may be defined by the general epistemic operator and the dependency operator, i.e. $L_{i,j}\varphi \stackrel{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$, which means that if i depends on j about φ and j has the belief φ, then i has sub-belief φ (from j). Therefore, under this definition about sub-beliefs, we commit ourselves to an interpretation where communications between agents are reliable, and every informant is honest. Moreover, we may view the general epistemic operator as

[3]Because we consider the axiom $D_{i,j}\varphi \leftrightarrow D_{i,j}\neg\varphi$ as a fundamental axiom about $D_{i,j}$

a kind of special sub-epistemic operator, i.e. $L_i\varphi \stackrel{\text{def}}{\Longleftrightarrow} L_{i,i}\varphi$. Therefore, we need sub-languages for belief dependence logics. We define the language $\mathbf{L_D}$ as the minimal set of formulas closed by the syntactic rules (i), (ii), (iii), (iv), and (vi). Furthermore, the language $\mathbf{L_L}$ is defined by the rules (i), (ii), (iii), (iv), the language $\mathbf{L_{Lij}}$ is defined by the rules (i), (ii), (iii), (v), and the language $\mathbf{L_{LijD}}$ is defined by the rule (i), (ii), (iii), (v), and (vi).

2.5 General Scenario

We have argued that an appropriate procedure for formalizing information assimilation should consist of two phases: compartmentalized and incorporated information. In the logics for belief dependence, compartmentalized information for agent i corresponds to sub-beliefs $L_{i,j}\varphi$, whereas incorporated information corresponds to general beliefs of agent i, namely $L_i\varphi$.

Cognitive psychology has presently not yet offered an available theory which specifies the process how and when an rational agent transfers her compartment beliefs into her incorporated beliefs. Sometimes the process depends on the agent's own belief maintenance strategy, which suggests the possibility to formalize the second phase using some theory of belief revision, which has been one of hot topics in the researches of AI and epistemology in the past years [Gärdenfors 1988, Gärdenfors 1990, Gärdenfors&Makinson 1988, Huang 1991b, Martin&Shapiro 1986, Nebel 1990].

In multi-agent environments, it is assumed that some primitive rely-on relations (about some propositions) among agents have been decided at the metalevel. This assumption is called *initial role-information assumption*. I believe that the assumption is appropriate and intuitive because, in multi-agent environments, some agents must have some *minimal* information about others to guarantee that communication. In many cases, primitive rely-on relations are easy to determine, because they can be viewed as those relations that are independent of the problem of belief updates. In a reliable communication network, if it is assumed that agents are both honest and unsuspicions, primitive rely-on relations often collapse into primitive communication relations. Note that this assumption does not require that each agent must have a complete knowledge about other agents. Actually initial role information may be updated after their communication.

Therefore, based on the primitive rely-on relations, we can obtain a complete knowledge about agents' sub-beliefs by using the logics for belief dependence. Furthermore, based on the complete sets concerning agents' sub-beliefs, some agents' appraisal of other agents may be determined. In the next chapter, some role-appraisal axioms such as "cautious believer", and "stubborn believer", are proposed. Based on these role-appraisal information, we can ultimately determine some rational belief maintenance strategies. However, as mentioned above, I first concentrate on the formalism concerning the first phase of information assimilation, i.e., on the problem of how the complete sub-belief and the complete rely-on relations can be determineded, based on the primitive rely-on relations. Then, I will move to the second phase of the formalization. The general scenario about the formalism of belief dependence is

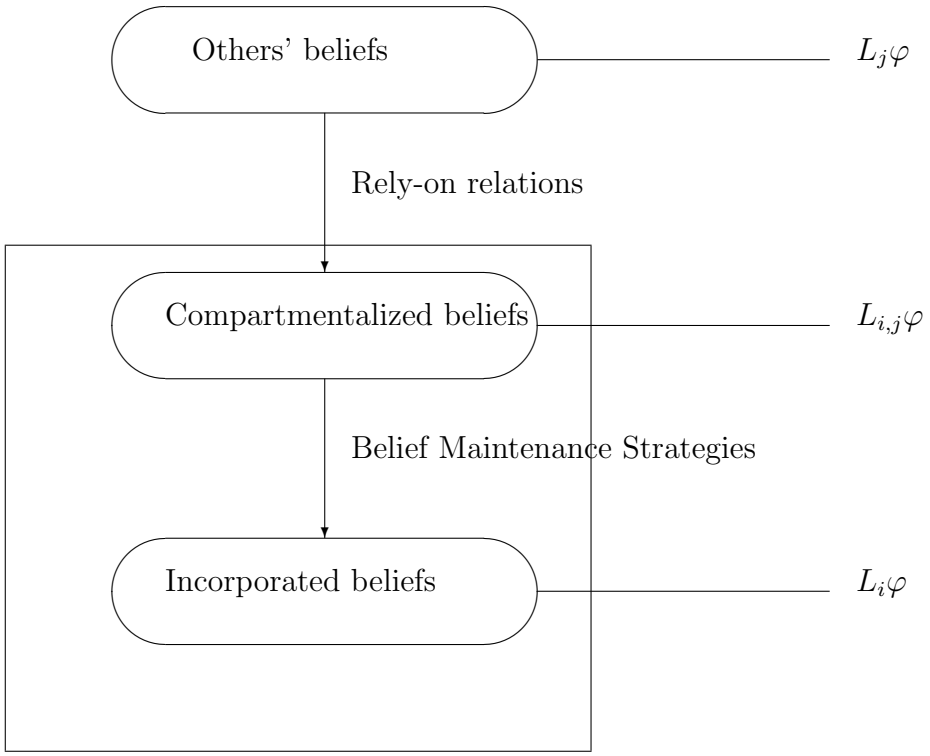


Figure 2.5: General Scenario

shown in the figure 2.5.

Chapter 3

Formalizing Belief Dependence

3.1 Several Plausible Systems

There are two approaches to logics. The first is a normal one from semantic models to characterizations, namely, first we try to capture an intuitive semantic model, and next we try to axiomize it. The second is somewhat less standard, from characterizations to semantic models. Under this second approach, we first have a characterization on the problem which we are concerned with, and next we try to invent some corresponding semantic model. In the formalization of the problem of belief dependence, notwithstanding the fact that the first approach is the more ideal one, there are difficulties to capture the semantic models for believed dependence without first considering their characterization. Because, I feel that a really intuitive and powerful semantic model for the belief dependence logics must reflect the details of communication system on which agents' activities are based, including the details of agents' psychological states, this will make the semantic models extremely complicated, and even elusive. Therefore, we use the second approach for formalizing the belief dependence. Moreover, I do not require that the interpretation of the belief dependence operators has a unique meaning. Each operator may have several different meanings. For instance, $L_{i,j}\varphi$ may be interpreted as " agent i relies on agent j for φ", or alternatively, as "agent j is the credible advisor of agent i about φ". I leave the decision of the exact meaning postulates for the operators to the user. Users can use the language of the logic of belief dependence to syntactically characterize their understanding of the meaning of the operators. Consequently, there exist many different logical systems for the belief dependence. In this section, I present a number of possible systems.

3.1.1 Belief Dependence Systems Based on the Epistemic Operator and the Dependency Operator

In this subsection, we define several belief dependence systems in which the sub-belief operator is defined in term of the epistemic operator and dependency operator. So

the corresponding language is $\mathbf{L}_D$. A natural way to achieve this by is defining $L_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$. This definition implies the axiom $L_{i,j}\varphi \rightarrow L_j\varphi$, which says that informants in the system are honest, and the axiom $D_{i,j}\varphi \wedge L_j\varphi \rightarrow L_{i,j}\varphi$, which means that the communication in this system is reliable. The resulting minimal system based on the epistemic operator and the dependency operator, is called **LD** system.

Axioms:
(BA) All instances of propositional tautologies.
(KL) $L_i\varphi \wedge L_i(\varphi \rightarrow \psi) \rightarrow L_i\psi$.

Rules of Inference:
(MP) $\vdash \varphi, \vdash \varphi \rightarrow \psi \Rightarrow \vdash \psi$.
(NECL) $\vdash \varphi \Rightarrow \vdash L_i\varphi$.

Definitions:
(Lijdf) $L_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$.

We may add additional axioms to this minimal system. Naturally there are various ways to do so. One of the possibilities is to include the weak-S5 system (for epistemic operator) as a subsystem of the belief dependence system. The resulting system for belief dependence, is called the $\mathbf{L5}^{-}$+**D4** system:

Axioms:
(BA) All instances of propositional tautologies.
(KL) $L_i\varphi \wedge L_i(\varphi \rightarrow \psi) \rightarrow L_i\psi$.
(NL) $\neg L_i\bot$.
(4L) $L_i\varphi \rightarrow L_iL_i\varphi$.
(5L) $\neg L_i\varphi \rightarrow L_i\neg L_i\varphi$.

The axioms above consist of a weak-S5 modal logic system. We add the following axioms about the dependency operator:

(D$\neg$) $D_{i,j}\varphi \leftrightarrow D_{i,j}\neg\varphi$.
(Neutrality axiom. Rely on someone else about φ iff rely on her about the negation of φ, this seems to be the most fundamental axiom for the dependency operator.)

(D$\rightarrow$) $D_{i,j}\varphi \wedge D_{i,j}(\varphi \rightarrow \psi) \rightarrow D_{i,j}\psi$.
(Closure under implication for the dependency operator, closing under implication seems to be a reasonable condition.)

(D$\wedge$) $D_{i,j}\varphi \wedge D_{i,j}\psi \rightarrow D_{i,j}(\varphi \wedge \psi)$.
(Closure under conjunction. Because we index sub-beliefs by agent name, this entails that beliefs which come from the same agent should be consistent.)

(DIPL) $D_{i,j}\varphi \rightarrow L_i D_{i,j}\varphi$.
(Positively explicit dependency axiom. As it is argued above, this axiom means that dependencies are explicitly known by the believer.)

Rules of Inference:
(MP) $\vdash \varphi, \vdash \varphi \rightarrow \psi \Rightarrow \vdash \psi$.
(NECL) $\vdash \varphi \Rightarrow \vdash L_i\varphi$.

Definition:
(Lijdf) $L_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$.

Remarks: By the neutrality axiom and the closure under conjunction axiom, we have, $\vdash D_{i,j}\varphi \Rightarrow \vdash D_{i,j}\varphi \wedge D_{i,j}\neg\varphi \Rightarrow \vdash D_{i,j}\bot$. Thus, we have the theorem, $D_{i,j}\varphi \rightarrow D_{i,j}\bot$, which means that if the agent i relies on the agent j about any proposition φ, then the agent i also relies on the agent j about the falsum. Belief dependence on the falsum seems to be counterintuitive. However, this does not cause a problem, since we define $L_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$ and we have $\neg L_j\bot$. So we still have the property $\neg L_{i,j}\bot$.

3.1.2 Belief Dependence System Based on Sub-belief Operator

Based on the sub-belief operator, we also can present a minimal logic for this approach, called the **LIJ** system,[1] which consists of the following axioms, inference rules, and definitions:

Axioms
(BA) All instances of propositional tautologies.

(KLij) $L_{i,j}\varphi \wedge L_{i,j}(\varphi \rightarrow \psi) \rightarrow L_{i,j}\psi$.
(Just as in general epistemic logics, sub-beliefs are closed under logical implication.)

Rules of Inference
(MP) $\vdash \varphi, \vdash \varphi \rightarrow \psi \Rightarrow \vdash \psi$.

(NECLij) $\vdash \varphi \Rightarrow \vdash L_{i,j}\varphi$.

Definitions
(Ddf) $D_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} L_{i,j}\varphi \vee L_{i,j}\neg\varphi$.
(If agent i believes φ or believes $\neg\varphi$ from agent j, then that agent i relies on agent j about φ.)

[1] the corresponding language is $\mathbf{L_{Lij}}$.

(Ldf) $L_i\varphi \overset{\text{def}}{\Longleftrightarrow} L_{i,i}\varphi$.
(We make no distinction between implicit beliefs and explicit beliefs.)

We also can enlarge this minimal system into other systems. The system called **Lij5**$^-$+**D**, which is a system by adding the following axioms into **LIJ** system:

(DLij) $\neg L_{i,j}\bot$.
(This axiom means that an agent never believes a false fact.)

(4Lij) $L_{i,j}\varphi \rightarrow L_iL_{i,j}\varphi$.
(Positive introspective axiom for sub-beliefs.)

(5Lij) $\neg L_{i,j}\varphi \rightarrow L_i\neg L_{i,j}\varphi$.
(Negative introspective axiom for sub-beliefs.)

The relation with the system **L5**$^-$+**D4** is expressed by the following:

3.1.1. Proposition. *Axioms* $(KL), (NL), (4L), (5L), (D\neg), (D\wedge), (DIPL)$ *are provable from the logic system* **Lij5**$^-$+**D**.

PROOF. (a) The axioms concerning the modal operator L_i, namely, axioms (KL)-(5L), directly follow from their corresponding axioms in the logic system **Lij5**$^-$+**D**, because the modal operator $L_{i,j}$ subsumes the modal operator $L_{i,i}$, and $L_{i,i}$ is equal to the modal operator L_i by the definition (Ldf).

(b) According to the definition (Ddf), the neutrality axiom (D$\neg$),
$D_{i,j}\varphi \rightarrow D_{i,j}\neg\varphi$, is evident, because $\vdash D_{i,j}\varphi \Rightarrow\vdash L_{i,j}\varphi \vee L_{i,j}\neg\varphi \Rightarrow\vdash L_{i,j}\neg\varphi \vee L_{i,j}\neg(\neg\varphi) \Rightarrow\vdash D_{i,j}\neg\varphi$.

(c) Closure under conjunction,
(D$\wedge$) $D_{i,j}\varphi \wedge D_{i,j}\psi \rightarrow D_{i,j}(\varphi \wedge \psi)$.
$\vdash D_{i,j}\varphi \wedge D_{i,j}\psi \Rightarrow\vdash (L_{i,j}\varphi \vee L_{i,j}\neg\varphi) \wedge (L_{i,j}\psi \vee L_{i,j}\neg\psi)$
$\Rightarrow\vdash L_{i,j}(\varphi \wedge \psi) \vee L_{i,j}(\neg\varphi \wedge \psi) \vee L_{i,j}(\varphi \wedge \neg\psi) \vee L_{i,j}(\neg\varphi \wedge \neg\psi)$
$\Rightarrow\vdash L_{i,j}(\varphi \wedge \psi) \vee L_{i,j}\neg(\varphi \wedge \psi)$
(Because $\vdash L_{i,j}(\neg\varphi \wedge \psi) \rightarrow L_{i,j}\neg(\varphi \wedge \psi)$, $\vdash L_{i,j}(\varphi \wedge \neg\psi) \rightarrow L_{i,j}\neg(\varphi \wedge \psi)$, and $\vdash L_{i,j}(\neg\varphi \wedge \neg\psi) \rightarrow L_{i,j}\neg(\varphi \wedge \psi)$)
$\Rightarrow\vdash D_{i,j}(\varphi \wedge \psi)$.

(d) Positively explicit dependence, namely,
(DIPL) $D_{i,j}\varphi \rightarrow L_iD_{i,j}\varphi$.
$\vdash D_{i,j}\varphi \Rightarrow\vdash L_{i,j}\varphi \vee L_{i,j}\neg\varphi \Rightarrow\vdash L_{i,i}L_{i,j}\varphi \vee L_{i,i}L_{i,j}\neg\varphi$
$\Rightarrow\vdash L_{i,i}(L_{i,j}\varphi \vee L_{i,j}\neg\varphi) \Rightarrow\vdash L_iD_{i,j}\varphi$. □

3.2 Formalizing Suspicion and Other Features

Based on the three modal operators concerning belief dependence, i.e., the general epistemic operator L_i, the sub-belief operator $L_{i,j}$, and the dependency operator $D_{i,j}$, we can formalize important and interesting features of belief dependence. The following axioms are candidates for formalizing properties of belief dependence.

(a) No-doubt Axiom
$L_{i,j}\varphi \rightarrow L_iL_j\varphi$.
(Whatever comes from someone else is believed to be a true belief of the others. We know that $L_{i,j}\varphi$ is not necessarily equal to $L_iL_j\varphi$. However, in the no-doubt belief dependence system, the sub-belief $L_{i,j}\varphi$ implies $L_iL_j\varphi$.)

(b) Honesty Axiom
$L_{i,j}\varphi \rightarrow L_j\varphi$.
(Sub-beliefs are believed by their informants. If we apply the definition $L_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$, then this means that every agent is honest.)

(c) Confidence Axiom
$L_i\varphi \wedge D_{i,j}\varphi \rightarrow L_iL_j\varphi$.
(Agent i believes her dependent beliefs are actually true beliefs of the agent she depends on.)

(d) Fool Believer Axiom
$L_i\varphi \rightarrow \exists j L_{i,j}\varphi, \quad (j \neq i)$[2].
(All of the agents' beliefs come from someone else.)

(e) Stubborn Believer Axiom
$L_{i,j}\varphi \rightarrow L_i\varphi$.
(The agent never believes somebody else's beliefs.)

(f) Communicative Agent Axiom
$L_i\varphi \rightarrow \exists j L_{j,i}\varphi, \quad (j \neq i)$.
(All of the agents' beliefs are believed by someone else.)

(g) Cautious Believer Axiom
$L_{i,j}\varphi \rightarrow \exists k L_{i,k}\varphi, \quad (k \neq j)$.
(The agent believes only those propositions which are believed by at least two agents.)

(h) Decomposition under Conjunction Axiom (D$\wedge'$)
$D_{i,j}(\varphi \wedge \psi) \rightarrow D_{i,j}\varphi \wedge D_{i,j}\psi$.

[2]Although we do not introduce any quantifier or equality in the language **L**, because we generally consider a finite agent set, say $\mathbf{A_n} = \{i_1, ..., i_n\}$, the formula $\exists j L_{i_l,j}\varphi, \ (j \neq i_l)$ can be viewed as an abbreviation for the formula $L_{i_l,i_1}\varphi \vee ... \vee L_{i_l,i_{l-1}}\varphi \vee L_{i_l,i_{l+1}}\varphi \vee ... \vee L_{i_l,i_n}\varphi$.

(This is a stronger axiom than the closure under implication axiom (D→), because, the neutral axiom and the decomposition under conjunction axiom together imply the closure under implication axiom.)

The following twelve axioms express various possible schemes:

(i) Positively Explicit Dependency Axiom (DIPL)
$D_{i,j}\varphi \rightarrow L_i D_{i,j}\varphi$.
(The positive dependency is explicitly known by the believer.)

(j) Negatively Explicit Dependency Axiom (DINL)
$\neg D_{i,j}\varphi \rightarrow L_i \neg D_{i,j}\varphi$.
(The negative dependency is explicitly known by the believer.)

(k) Positively Explicit Dependency Axiom (DJPL)
$D_{i,j}\varphi \rightarrow L_j D_{i,j}\varphi$.
(The positive dependency is explicitly known by the relied agent.)

(l) Negatively Explicit Dependency Axiom (DJNL)
$\neg D_{i,j}\varphi \rightarrow L_j \neg D_{i,j}\varphi$.
(The negative dependency is explicitly known by the relied agent.)

(m) Positively Reliable Dependence Axiom (DIPR)
$L_i D_{i,j}\varphi \rightarrow D_{i,j}\varphi$.
(The agent's information about the positive dependency is correct.)

(n) Negative Reliable Dependence Axiom (DINR)
$L_i \neg D_{i,j}\varphi \rightarrow \neg D_{i,j}\varphi$.
(The agent's information about non-dependency is correct.)

(o) Positively Reliable Dependence Axiom (DJPR)
$L_j D_{i,j}\varphi \rightarrow D_{i,j}\varphi$.
(The relied agent's information about the positive dependency is correct.)

(p) Negative Reliable Dependence Axiom (DJNR)
$L_j \neg D_{i,j}\varphi \rightarrow \neg D_{i,j}\varphi$.
(The relied agent's information about the negative dependency is correct.)

(q) Axiom of Common Knowledge of Positive Dependence (DKPL)
$D_{i,j}\varphi \rightarrow L_k D_{i,j}\varphi$.
(Agents' positive dependencies are known by any other. The axiom subsumes axiom (DIPL) and axiom (DJPL).)

(r) Axiom of Common Knowledge of Negative Dependence (DKNL)

$\neg D_{i,j}\varphi \rightarrow L_k \neg D_{i,j}\varphi$.
(Agents' positive dependencies are known by any other. The axiom subsumes (DINL) and (DJNL).)

(s) Positive axiom of reliable dependence (DKPR)
$L_k D_{i,j}\varphi \rightarrow D_{i,j}\varphi$.
(Agents' information about the positive dependency is correct. The subsumption is clear.)

(t) Negative axiom of reliable dependence (DKNR)
$L_k \neg D_{i,j}\varphi \rightarrow \neg D_{i,j}\varphi$.
(Agents' information about negative dependency is correct. The subsumption is clear.)

The axioms (i) to (t), called *dependency introspection axioms*, are specially interesting, since those axioms characterize certain properties about the dependency information. These introspection axioms share a common syntactic structure which can be specified as follows. We define a specification function $DSformula : A_n \times A_n \times A_n \times \{1,0\} \times \{l,r\} \rightarrow \mathbf{L}$ as follows:

$$DSformula(i,j,k,x,y) = \begin{cases} D_{i,j}\varphi \rightarrow L_k D_{i,j}\varphi & if \quad x=1, y=l \\ \neg D_{i,j}\varphi \rightarrow L_k \neg D_{i,j}\varphi & if \quad x=0, y=l \\ L_k D_{i,j}\varphi \rightarrow D_{i,j}\varphi & if \quad x=1, y=r \\ L_k \neg D_{i,j}\varphi \rightarrow \neg D_{i,j}\varphi & if \quad x=0, y=r \end{cases}$$

Therefore, $DSformula(i,j,k,1,l)$ = axiom (DKPL), $DSformula(i,j,i,1,l)$ = axiom (DIPL), $DSformula(i,j,j,1,l)$ =axiom (DJPL). Other relations can be similarly obtained.

Moreover, based on those operators, we can formalize the notion of suspicion as follows:
$\text{Suspect}_i\varphi \overset{\text{def}}{\Longleftrightarrow} (\exists j)(L_{i,j}\varphi \wedge \neg L_i L_j \varphi)$.
(Agent i suspects φ if and only if there exists some agent j such that agent i believes φ on behalf of j, but agent i does not believe that agent j believes φ.)

3.2.1. Proposition. *For the system* **Lij5**$^-$**+D**:
(a) $Suspect_i\varphi \rightarrow L_i Suspect_i\varphi$.
(If agent i suspects φ, then he knows that fact.)

(b) $\neg Suspect_i\varphi \rightarrow L_i \neg Suspect_i\varphi$.
(If agent i does not suspect φ, then he knows that fact.)

PROOF. (a) $\vdash \text{Suspect}_i\varphi \Leftrightarrow \vdash (\exists j)(L_{i,j}\varphi \wedge \neg L_i L_j \varphi)$
$\Rightarrow \vdash L_{i,j}\varphi \wedge \neg L_i L_j \varphi \Rightarrow \vdash L_i L_{i,j}\varphi \wedge L_i \neg L_i L_j \varphi$
$\Rightarrow \vdash L_i(L_{i,j}\varphi \wedge \neg L_i L_j \varphi) \Rightarrow \vdash L_i \text{Suspect}_i\varphi$.

(b)$\vdash \neg\text{Suspect}_i\varphi \Leftrightarrow\vdash (\forall j)(\neg L_{i,j}\varphi \vee L_iL_j\varphi)$
$\Rightarrow\vdash (\forall j)(L_i\neg L_{i,j}\varphi \vee L_iL_iL_j\varphi) \Rightarrow\vdash L_i((\forall j)(\neg L_{i,j}\varphi \vee L_iL_j\varphi))$
$\Rightarrow\vdash L_i\neg\text{Suspect}_i\varphi$. □

3.3 Formalizing Indirect Dependence

In multiple agent environments, beliefs may be transitive among agents. Therefore, we extend the definition of dependent beliefs to indirect dependent beliefs as follows:

(D+df) $D^+_{i,j}\varphi \stackrel{\text{def}}{\Longleftrightarrow} D_{i,j_1}\varphi \wedge D_{j_1,j_2}\varphi \wedge ... \wedge D_{j_m,j}\varphi,\ (i \neq j_1)$,

(D∗df) $D^*_{i,j}\varphi \stackrel{\text{def}}{\Longleftrightarrow} D^+_{i,j}\varphi \vee D_{i,j}\varphi$.

We have the following propositions:

3.3.1. Proposition. *(Transitivity of Indirect Dependence)*
(a) $D^*_{i,j}\varphi \wedge D^*_{j,k}\varphi \rightarrow D^*_{i,k}\varphi$.
(b) $D^+_{i,j}\varphi \wedge D^+_{j,k}\varphi \rightarrow D^+_{i,k}\varphi$.

More generally, we have:
(c) for any $x, y, z \in \{*, +\},\ (i \neq j), (j \neq k)$,
$D^x_{i,j}\varphi \wedge D^y_{j,k}\varphi \rightarrow D^z_{i,k}\varphi$.
PROOF. *They are straightforward from the definition.* □

We also define indirect sub-beliefs for the agent set $\mathbf{A}_n$ as follows:
(Lij1df) $L^1_{i,j}\varphi \stackrel{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$.
(Lijmdf) $L^m_{i,j}\varphi \stackrel{\text{def}}{\Longleftrightarrow} D_{i,j'}\varphi \wedge L^{m-1}_{j',j}\varphi$.
(Lij∗df) $L^*_{i,j}\varphi \stackrel{\text{def}}{\Longleftrightarrow} [\vee^n_{k=1}]L^k_{i,j}\varphi$.

From the definitions above, we can easily show that the following propositions hold in any logic of belief dependence which consists of the axioms (BA), (KL), (DL), inference rules (MP), (NECL), and the definition (Lijdf) as its subsystem.

3.3.2. Proposition.
(a) Coincidence
$L^*_{i,j}\varphi \leftrightarrow D^*_{i,j}\varphi \wedge L_j\varphi$.
(b) Consistence
$L^*_{i,j}\varphi \rightarrow \neg L^*_{i,j}\neg\varphi$.
(c) Same-source-propagation

$D^*_{i,k}\varphi \wedge L^*_{j,k}\varphi \rightarrow L^*_{i,k}\varphi$.
(d) Strong-consistency
$L^*_{i,j}\neg\varphi \rightarrow (\neg L^*_{k,j}\varphi)$.
(e) No-same-source-assertion
$L^*_{i,j}\varphi \wedge \neg L^*_{k,j}\varphi \rightarrow \neg D^*_{k,j}\varphi$.

PROOF.
(a) $L^*_{i,j}\varphi \leftrightarrow D^*_{i,j}\varphi \wedge L_j\varphi$.
It is straightforward from the definition.

(b) $L^*_{i,j}\varphi \rightarrow \neg L^*_{i,j}\neg\varphi$.

$\vdash L^*_{i,j}\varphi \Rightarrow \vdash D^*_{i,j}\varphi \wedge L_j\varphi$
$\Rightarrow \vdash L_j\varphi \Rightarrow \vdash \neg L_j\neg\varphi$
$\Rightarrow \vdash \neg L_j\neg\varphi \vee \neg D^*_{i,j}\neg\varphi \Rightarrow \vdash \neg L^*_{i,j}\neg\varphi$.

(c) $D^*_{i,k}\varphi \wedge L^*_{j,k}\varphi \rightarrow L^*_{i,k}\varphi$.

$\vdash D^*_{i,k} \wedge L^*_{j,k}\varphi \Rightarrow \vdash D^*_{i,k}\varphi \wedge D^*_{j,k}\varphi \wedge L_k\varphi$ (By (a))
$\Rightarrow \vdash D^*_{i,k}\varphi \wedge L_k\varphi$
$\Rightarrow \vdash L^*_{i,k}\varphi$ (By (a))

(d) $L^*_{i,j}\neg\varphi \rightarrow (\neg L^*_{k,j}\varphi)$.
It is similar to that of (b).

$\vdash L^*_{i,j}\varphi \Rightarrow \vdash L_j\varphi \Rightarrow \vdash \neg L_j\neg\varphi$
$\Rightarrow \vdash \neg L_j\neg\varphi \vee \neg D^*_{k,j}\neg\varphi \Rightarrow \vdash \neg L^*_{k,j}\neg\varphi$.

(e) $L^*_{i,j}\varphi \wedge \neg L^*_{k,j}\varphi \rightarrow \neg D^*_{k,j}\varphi$.

$\vdash L^*_{i,j}\varphi \wedge \neg L^*_{k,j}\varphi \Rightarrow \vdash D^*_{i,j}\varphi \wedge L_j\varphi \wedge (\neg D^*_{k,j}\varphi \vee \neg L_j\varphi)$
$\Rightarrow \vdash D^*_{i,j}\varphi \wedge L_j\varphi \wedge \neg D^*_{k,j}\varphi \Rightarrow \vdash \neg D^*_{k,j}\varphi$. □

Chapter 4

Semantic Models

4.1 L-Model of Belief Dependence: an approach based on general epistemic logic

In this section, we first try to define the dependency operator by a reduction to general doxastic and epistemic operators, so that we can study the problem of belief dependence in the framework of the standard doxastic and epistemic logics. $D_{i,j}\varphi$ means that agent i relies on agent j about believing φ. Formally, there might exist many different interpretations of the dependency operator. In other words, there are many semantic interpretations for the notion of "rely on".

Here are some of the possible definitions:

(Ddf1) $D_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} (L_j\varphi \rightarrow L_i\varphi) \wedge (L_j\neg\varphi \rightarrow L_i\neg\varphi)$.
(If agent j believes φ, so does agent i; if agent j believes φ is false, agent i believes φ is false as well.)

(Ddf1') $D_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} (L_j\varphi \leftrightarrow L_i\varphi)$.
(If agent j believes φ, so does agent i; if agent j does not believe φ, neither does agent i)

(Ddf2) $D_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} L_i(L_j\varphi \rightarrow L_i\varphi) \wedge L_i(L_j\neg\varphi \rightarrow L_i\neg\varphi)$.
(Agent i believes that if agent j believes φ, then so does agent i, agent j believes its false, so does agent i.)

(Ddf2') $D_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} L_i(L_j\varphi \leftrightarrow L_i\varphi)$.
(Agent i believes that agent j believes φ iff agent i believes φ).

(Ddf3) $D_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} (L_iL_j\varphi \rightarrow L_i\varphi) \wedge (L_iL_j\neg\varphi \rightarrow L_i\neg\varphi)$.
(If agent i believes that agent j believes φ, then agent i will believe it; if agent i believes agent j believes φ is false, then agent i will also believe that φ is false.)

Of those definitions, (Ddf2) and (Ddf2') are the definitions of explicit dependence, because they say that agent i believes the dependency relation, whereas other definitions are implicit. Moreover, (Ddf1) seems to be simple, but it is completely implicit. (Ddf3) is semi-implicit since agent i's dependent beliefs depend on parts of her own beliefs. (Ddf1') is a symmetric definition. However, dependent relations are not intuitively symmetric. Although (Ddf2') is not symmetric, the equivalence still makes the definition too strong. Therefore, we view the definitions (Ddf1), (Ddf2), and (Ddf3) as more reasonable and acceptable.

For those three definitions (Ddf1), (Ddf2), and (Ddf3), we know that the neutrality axiom (D1), namely, $D_{i,j}\varphi \leftrightarrow D_{i,j}\neg\varphi$, holds in any epistemic logic. Moreover, we intuitively expect that the closure under conjunction axiom should hold for those definitions. Unfortunely, we have the following result.

4.1.1. Claim. *For any of the standard Kripke style semantics for the epistemic operator L_i, $D_{i,j}\varphi \wedge D_{i,j}\psi \wedge \neg D_{i,j}(\varphi \wedge \psi)$ is satisfiable when $D_{i,j}\varphi$ is defined by either (Ddf1), (Ddf2), or (Ddf3).*

PROOF. For the definition (Ddf1),
$D_{i,j}\varphi \wedge D_{i,j}\psi \wedge \neg D_{i,j}(\varphi \wedge \psi)$
$\equiv (L_j\varphi \rightarrow L_i\varphi) \wedge (L_j\neg\varphi \rightarrow L_i\neg\varphi) \wedge (L_j\psi \rightarrow L_i\psi) \wedge (L_j\neg\psi \rightarrow L_i\neg\psi)$
$\wedge \neg((L_j(\varphi \wedge \psi) \rightarrow L_i(\varphi \wedge \psi)) \wedge (L_j\neg(\varphi \wedge \psi) \rightarrow L_i\neg(\varphi \wedge \psi)))$
$\equiv (L_j\varphi \rightarrow L_i\varphi) \wedge (L_j\neg\varphi \rightarrow L_i\neg\varphi) \wedge (L_j\psi \rightarrow L_i\psi) \wedge (L_j\neg\varphi \rightarrow L_i\neg\psi)$
$\wedge ((L_j(\varphi \wedge \psi) \wedge \neg L_i(\varphi \wedge \psi) \vee L_j(\neg\varphi \vee \neg\psi) \wedge \neg L_i(\neg\varphi \vee \neg\psi))$ (Formula 1)
Moreover, let (Formula 2) be the formula $\neg L_j\varphi \wedge \neg L_j\neg\varphi \wedge \neg L_j\psi \wedge \neg L_j\neg\psi \wedge \neg L_i(\neg\varphi \vee \neg\psi) \wedge L_j(\neg\varphi \vee \neg\psi)$.
We know that if (Formula 2) is satisfiable, then so is (Formula 1), because we have:
$\neg L_j\varphi \Rightarrow (L_j\varphi \rightarrow L_i\varphi)$
$\neg L_j\neg\varphi \Rightarrow (L_j\neg\varphi \rightarrow L_i\neg\varphi)$
$\neg L_j\psi \Rightarrow L_j\psi \rightarrow L_i\psi$
$\neg L_j\neg\psi \Rightarrow L_j\neg\psi \rightarrow L_i\psi$
$L_j(\neg\varphi \vee \neg\psi) \wedge \neg L_i(\neg\varphi \vee \neg\psi) \Rightarrow L_j(\neg\varphi \vee \neg\psi) \wedge \neg L_i(\neg\varphi \vee \neg\psi)$.
It is easy to show that (Formula 2) is satisfiable. One of the cases is shown in the figure. The cases of (Ddf2) and (Ddf3) can be shown similarly. □

The above argument illustrates that general epistemic logic is not an appropriate tool for formalizing the problem of belief dependence, since some intuitive properties such as closure under conjunction come out not to be valid. To make a comparison with other semantic models, we call these semantic models *belief dependence L-model*. Therefore, we have the following formal definition.

4.1.1. Definition. (Belief Dependence L-model) *A belief dependence L-model is a tuple $M = \langle W, \mathcal{L}, V\rangle$*
where W is a set of states, V is a truth assignment to each primitive proposition p in Φ_0 a set of possible worlds, and $\mathcal{L}$ is a function which consists of n binary accessibility relations on W, i.e., $\mathcal{L} : \mathbf{A_n} \rightarrow \mathcal{P}(W \times W)$.

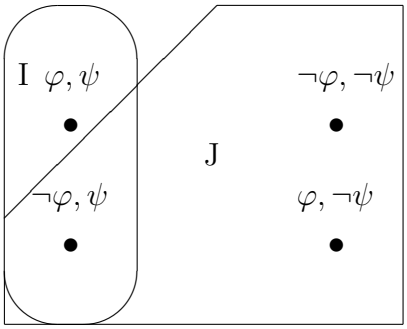


Figure 4.1: Satisfiability for (Formula 2)

4.2 D-Model of Belief Dependence: a syntactic approach

4.2.1 Semantics

We know that sub-beliefs can be defined directly in terms of the dependency operator and the general epistemic operator, namely, $L_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$. Therefore, to formulate belief dependence natural, a possible approach is to add a dependency structure to the standard Kripke model of epistemic logic. This approach is similar to Fagin and Halpern's general awareness logic[Fagin&Halpern 1988]. The idea is that one can decide on a metalevel for which formulas each agent is supposed to rely on others. By this approach, we introduce *dependency formula sets* for each agent pair i and j, and each possible world w, namely, formula sets $\mathcal{D}(i, j, w)$. The formula $\varphi \in \mathcal{D}(i, j, w)$ means that in world w agent i relies on agent j about the formula φ. Therefore, it is called a syntactic approach. Formally, we have the following definition:

4.2.1. Definition. (Belief dependence D-model) *A belief dependence D-model is a tuple* $M = \langle W, \mathcal{L}, \mathcal{D}, V\rangle$
where W *is a set of possible worlds,* V*, as usual, is a truth assignment function to each primitive proposition* $p \in \Phi_0$ *a subset of possible worlds, and* $\mathcal{L} : \mathbf{A_n} \to \mathcal{P}(W \times W)$*, which consists of* n *binary accessibility relations on* W*,* $\mathcal{D} : \mathbf{A_n} \times \mathbf{A_n} \times W \to \mathcal{P}(\mathbf{L_D})$ *is a dependency function.*

The truth relation $\models$ is defined inductively as follows:

$M, w \models p$,	iff $w \in V(p)$, where p is a primitive proposition,
$M, w \models \neg\varphi$	iff $M, w \not\models \varphi$,
$M, w \models \varphi_1 \wedge \varphi_2$	iff $M, w \models \varphi_1 \wedge M, w \models \varphi_2$,
$M, w \models L_i\varphi$	iff $M, t \models \varphi$ for all t such $\langle w, t\rangle \in \mathcal{L}(i)$,
$M, w \models D_{i,j}\varphi$	iff $\varphi \in \mathcal{D}(i, j, w)$.

We say a formula φ is *valid in structure* M, written $M \models \varphi$, if $M, w \models \varphi$ for all possible worlds w in M; φ is *satisfiable in* M if $M, w \models \varphi$ for some possible worlds

in M. We say φ is *valid* if it is valid in all structures; φ is *satisfiable* if it is satisfiable in some structure. A *belief dependence D-frame* is a tuple $\mathcal{F} = \langle W, \mathcal{L}, \mathcal{D} \rangle$ where W is a set of possible worlds, $\mathcal{L}$ consists of n accessibility relations on W, and $\mathcal{D}$ is a dependency function. We say a formula φ is *true in a frame* $\mathcal{F}$, written $\mathcal{F} \models \varphi$, if $\langle W, \mathcal{L}, \mathcal{D}, V \rangle, w \models \varphi$ for any truth valuation function V and any world $w \in W$.

For D-models, we define sub-beliefs as $L_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$, consequently the system is honest because the honesty axiom $L_{i,j}\varphi \rightarrow L_j\varphi$ holds.

In the definition of the D-model of belief dependence, we have placed no restrictions on the dependency formula sets. To capture certain properties for belief dependence, we add some restrictions on the dependency formula sets. Some typical restrictions we may want to add to $\mathcal{D}(i, j, w)$ can be expressed by some closure properties under the logical connectives and modal operators.

4.2.2. Definition. *A dependency formula set $\mathcal{D}(i, j, w)$ is said to be:*
(a) closed under negation, iff $\varphi \in \mathcal{D}(i, j, w) \Leftrightarrow \neg\varphi \in \mathcal{D}(i, j, w)$.
(b) closed under conjunction, iff $\varphi \in \mathcal{D}(i, j, w)$ and $\psi \in \mathcal{D}(i, j, w) \Rightarrow (\varphi \wedge \psi) \in \mathcal{D}(i, j, w)$.
(c) decomposable under conjunction, iff $\varphi \wedge \psi \in \mathcal{D}(i, j, w) \Rightarrow \varphi, \psi \in \mathcal{D}(i, j, w)$.
(d) closed under implication, iff $\varphi \in \mathcal{D}(i, j, w)$ and $(\varphi \rightarrow \psi) \in \mathcal{D}(i, j, w) \Rightarrow \psi \in \mathcal{D}(i, j, w)$.

It is easy to see that the above conditions (a), (b), (c), (d) correspond the axioms (D$\neg$), (D$\wedge$), (D$\wedge'$), and (D$\rightarrow$) respectively. At this place I won't to give the semantics conditions for all of the plausible axioms. However, the semantics conditions for the dependency properties axioms are specially interesting. Just like we deal with their counterparts in the syntax, we also introduce a frame specification function $DSframe : A_n \times A_n \times A_n \times \{1, 0\} \times \{l, r\} \rightarrow \mathcal{P}(Dframe)$ where $Dframe$ is the set of all D-frames.

4.2.3. Definition. *For any agent $i, j, k \in A_n$,*
(a) $DSframe(i, j, k, 1, l)$ is the set of all D-frames $\mathcal{F} = \langle W, \mathcal{L}, \mathcal{D} \rangle$ which satisfy the following condition:
for any formula φ, and any possible world w, if $\varphi \in \mathcal{D}(i, j, w)$, then $\varphi \in \mathcal{D}(i, j, t)$ for all of possible worlds t such that $\langle w, t \rangle \in \mathcal{L}(k)$.

(b) $DSframe(i, j, k, 0, l)$ is the set of all D-frames $\mathcal{F} = \langle W, \mathcal{L}, \mathcal{D} \rangle$ which satisfy the following condition:
for any formula φ, and any possible world w, if there exists a possible world t such that $\langle w, t \rangle \in \mathcal{L}(k)$ and $\varphi \in \mathcal{D}(i, j, t)$, then $\varphi \in \mathcal{D}(i, j, w)$.

(c) $DSframe(i, j, k, 1, r)$ is the set of all D-frames $\mathcal{F} = \langle W, \mathcal{L}, \mathcal{D} \rangle$ which satisfy the following condition:
for any formula φ, and any possible world w, if, for all $t \in W$ such that $\langle w, t \rangle \in \mathcal{L}(k), \varphi \in \mathcal{D}(i, j, t)$, then $\varphi \in \mathcal{D}(i, j, w)$.

(d) $DSframe(i,j,k,0,r)$ *is the set of all D-frames* $\mathcal{F} = \langle W, \mathcal{L}, \mathcal{D}\rangle$ *which satisfy the following condition:*
for any formula φ*, and any possible world* w*, if* $\varphi \in \mathcal{D}(i,j,w)$*, then there exists a possible world* t *such that* $\langle w,t\rangle \in \mathcal{L}(k)$ *and* $\varphi \in \mathcal{D}(i,j,t)$*.*

These frame properties correspond to the introspection axioms $DSformula(i,j,k,x,y)$ which are introduced in the chapter 3.

4.2.4. THEOREM. *For any* $i,j,k \in A_n$*, any* $x \in \{1,0\}$*, and any* $y \in \{l,r\}$*,* $DSformula(i,j,k,x,y)$ *is true in a D-frame* $\mathcal{F} = \langle W, \mathcal{L}, \mathcal{D}\rangle$ *iff* $\mathcal{F} \in DSframe(i,j,k,x,y)$*.*

PROOF. The proofs for the cases in which $x = 1$ and the cases in which $x = 0, y = r$ are straightforward. The only non-trivial case in the proof is the case $x = 0, y = l$. The corresponding axiom for $DSformula(i,j,k,0,l)$ is (DKNL), namely, $\neg D_{i,j}\varphi \rightarrow L_k \neg D_{i,j}\varphi$.

($\Rightarrow$) Suppose that $DSformula(i,j,k,0,l)$ is true in $\mathcal{F}$. Thus, for any valuation function V and any w, $M = \langle W, \mathcal{L}, \mathcal{D}, V\rangle, w \models \neg D_{i,j}\varphi \rightarrow L_k \neg D_{i,j}\varphi$. Moreover, if there exists a t such that $\langle w,t\rangle \in \mathcal{L}(k)$ and $\varphi \in \mathcal{D}(i,j,t)$, we have $M, w \models \neg L_k \neg D_{i,j}\varphi$. So $M, w \models D_{i,j}\varphi$. requiring $\varphi \in \mathcal{D}(i,j,w)$. Therefore, we conclude that $\mathcal{F} \in DSframe(i,j,k,0,l)$.

($\Leftarrow$) Suppose that $\mathcal{F} \in DSframe(i,j,k,0,l)$ and $\langle W, \mathcal{L}, \mathcal{D}, V\rangle, w \models \neg L_k \neg D_{i,j}\varphi$ for any V and w. By the truth condition, we have that there exists a t such that $\langle w,t\rangle \in \mathcal{L}(k)$ and $\varphi \in \mathcal{D}(i,j,t)$. So $\varphi \in \mathcal{D}(i,j,w)$ by the frame condition. Therefore, $\mathcal{F} \models \neg L_k \neg D_{i,j}\varphi \rightarrow D_{i,j}\varphi$. By contraposition, $\mathcal{F} \models \neg D_{i,j}\varphi \rightarrow L_k \neg D_{i,j}\varphi$. □

For the semantic models of the system $\mathbf{L5^-}$+**D4**, we have the following:

4.2.5. DEFINITION. *A D-model for belief dependence* $M = \langle W, \mathcal{L}, \mathcal{D}, V\rangle$ *is an* $\mathbf{L5^-}$+**D4** *D-model, if it satisfies the following conditions:*
(a) Each accessibility relation $\mathcal{L}(i)$ *is serial, transitive, and Euclidean,*
(b) Each dependency formula set $\mathcal{D}(i,j,w)$ *is closed under negation, implication, and conjunction,*
(c) The dependency function $\mathcal{D}$ *is of positive explicit dependency, namely, the corresponding frame* $\mathcal{F} \in DSframe(i,j,i,1,l)$*.*

4.2.2 Soundness and Completeness

In order to show soundness and completeness of the system $\mathbf{L5^-}$+**D4** for $\mathbf{L5^-}$+**D4** D-models, we use the standard techniques of building canonical structures[Huang 1989, Fagin&Halpern 1988, Hughes 1984]. First, we need the following definitions: A formula φ is *consistent* (with respect to an axiom system) if $\neg\varphi$ is not provable. A finite

set $\{\varphi_1, ..., \varphi_k\}$ is consistent iff the formula $\varphi_1 \wedge ... \wedge \varphi_k$ is consistent. An infinite set of formulae is consistent if every finite subset of it is consistent. A set F of formulae is a *maximally consistent set* if it is consistent and any strict superset is inconsistent. As it is pointed out in [Fagin&Halpern 1988], using standard techniques of propositional reasoning we can show:

4.2.6. LEMMA. (Maximal consistent set lemma) *In any axiom system that includes (BA) and (MP):*
(1) Any consistent set can be extended to a maximal consistent set.
(2) If F is a maximal consistent set, then for all formulas φ and ψ:
(2.a) either $\varphi \in F$ or $\neg\varphi \in F$,
(2.b) $\varphi \wedge \psi \in F$ iff $\varphi \in F$ and $\psi \in F$,
(2.c) if $\varphi \in F$ and $\vdash \varphi \rightarrow \psi$, then $\psi \in F$,
(2.d) if φ is provable, then $\varphi \in F$.

The completeness of a logic S means that

(A) For arbitrary formula set Δ and arbitrary formula ϕ, $\Delta \models \phi \Rightarrow \ \Delta \vdash_S \phi$

It actually turns out to be easier to show the following statement:

(B) For arbitrary formula set Δ, Δ is consistent with $S \ \ \Leftrightarrow \Delta$ has an S-model.

It can be easily shown that (A) and (B) are equivalent. Assume that we can construct a canonical model M_c where the possible worlds are maximal consistent sets, then, in order to show the completeness, we have to show that for any formula ϕ,

(1) $\phi \in w \Leftrightarrow M_c, w \models \phi$
(2) M_c is an $\mathbf{L5}^-$+**D4** D-model

We call the above (1) *the truth lemma.*

For any maximal consistent set w, we use $L_i^-(w)$ to denote the set $\{\varphi : L_i\varphi \in w\}$

4.2.7. LEMMA. *If a maximal consistent set w contains a formula $\neg L_i\varphi$, then $L_i^-(w) \cup \{\neg\varphi\}$ is consistent.*

PROOF. Suppose that w is a maximal consistent set such that $\neg L_i\varphi \in w$. We would like to show that if $L_i^-(w) \cup \{\neg\varphi\}$ is not consistent, then neither is w.
Suppose that $L_i^-(w) \cup \{\neg\varphi\}$ is not consistent. This means that there is some finite subset $\{\phi_1, ..., \phi_k\}$ of $L^-(w)$ such that $\neg(\phi_1 \wedge ... \wedge \phi_k \wedge \neg\varphi)$ is provable. Therefore, $(\phi_1 \wedge ... \wedge \phi_k) \rightarrow \varphi$ is provable. So $L_i(\phi_1 \wedge ... \wedge \phi_k) \rightarrow L_i\varphi$ is provable. Moreover, $(L_i\phi_1 \wedge ...L_i\phi_k) \rightarrow L_i\varphi$ is also provable. This means that $\{L_i\phi_1, ..., L_i\phi_k, \neg L_i\varphi\}$ is not consistent. Since it is a subset of w, w is not consistent. $\square$

4.2.8. THEOREM. **L5**$^-$+**D4** *belief dependence systems are sound and complete in the class of* **L5**$^-$+**D4** *D-models.*

Proof:. Soundness is evident. For the completeness, a canonical structure M_c is constructed as follows:
$M_c = \langle W, \mathcal{L}_c, \mathcal{D}_c, V\rangle$
where
$W = \{v : v \text{ is a maximal consistent set}\}$,
$v \in V(p)$ iff $p \in v$, for any primitive proposition p,
$\mathcal{L}_c(i) = \{\langle v, w\rangle : L_i^-(v) \subseteq w\}$, for any $i \in \mathbf{A_n}$,
$\mathcal{D}_c(i, j, v) = \{\varphi : D_{i,j}\varphi \in v\}$.

First, we show that M_c is an **L5**$^-$+**D4** D-model. Axioms (L3), (L4), and (L5) guarantee that $\mathcal{L}_c(i)$ is serial, transitive, and Euclidean. As far as the dependence function $\mathcal{D}_c$ is concerned, we have:

$\varphi \in \mathcal{D}_c(i, j, v) \Rightarrow D_{i,j}\varphi \in v$ (By the definition of M_c)
$\Rightarrow D_{i,j}\neg\varphi \in v$ (By axiom (D$\neg$) and lemma (2.c))
$\Rightarrow \neg\varphi \in \mathcal{D}_c(i, j, v)$ (By the definition of M_c)

Therefore, the dependence formula sets are closed under negation. The cases concerning closure under conjunction and implication can be similarly shown. Moreover, for any formula φ,

$\varphi \in \mathcal{D}_c(i, j, v) \Rightarrow D_{i,j}\varphi \in v$ (By the definition of M_c)
$\Rightarrow L_i D_{i,j}\varphi \in v$ (By axiom (DIPL) and lemma (2.c))
$\Rightarrow D_{i,j}\varphi \in W$ for all W such that $\langle v, w\rangle \in \mathcal{L}_c(i)$ (By the definition of M_c)
$\varphi \in \mathcal{D}_c(i, j, w)$ for all w such that $\langle v, w\rangle \in \mathcal{L}_c(i)$.
Therefore, M_c is an **L5**$^-$+**D4** D-model.

In order to show that every formula φ is satisfiable, we must show $\varphi \in v \Leftrightarrow M_c, v \models \varphi$. we can show this by induction on the structure of formulas as follows:

Case (a) φ is a primitive proposition, case (b) $\varphi \wedge \psi$, and case (c) $\neg\varphi$, are straightforward.

(d) $L_i\varphi$,
($\Leftarrow$) We would like to show that $L_i\varphi \notin v \Rightarrow M_c, v \not\models L_i\varphi$.

$L_i\varphi \notin v$
$\Rightarrow \neg L_i\varphi \in v$ (Maximal consistent set of v)
$\Rightarrow L_i^-(v) \cup \{\neg\varphi\}$ is consistent. (Lemma 4.2.7)
$\Rightarrow \exists w \in W(L_i^-(v) \cup \{\neg\varphi\} \subseteq w)$ (Maximal consistent set lemma (1))
$\Rightarrow \exists w \in W(L_i^-(v) \cup \{\neg\varphi\} \subseteq w$ and $\langle v, w\rangle \in \mathcal{L}_c(i))$ (Definition of $\mathcal{L}_c(i)$)
$\Rightarrow \exists w \in W(\langle v, w\rangle \in \mathcal{L}_c(i)$ and $\neg\varphi \in w)$ (Meta reasoning)

$\Rightarrow \exists w \in W(\langle v, w\rangle \in \mathcal{L}_c(i)$ and $\varphi \notin w)$ (Maximal consistent set)
$\Rightarrow \exists w \in W(\langle v, w\rangle \in \mathcal{L}_c(i)$ and $M_c, w \not\models \varphi)$ (Induction hypothesis)
$\Rightarrow M_c, v \not\models L_i\varphi$ (Truth condition)

$(\Rightarrow)$
$L_i\varphi \in v \Rightarrow \varphi \in w$ for all w such that $\langle v, w\rangle \in \mathcal{L}_c(i)$ (By the definition of M_c)
$\Leftrightarrow M_c, w \models \varphi$ for all w such that $\langle v, w\rangle \in \mathcal{L}_c(i)$ (Induction hypothesis)
$\Leftrightarrow M_c, v \models L_i\varphi$ (Truth condition)

(e) $D_{i,j}\varphi$,
$D_{i,j}\varphi \in v \Leftrightarrow \varphi \in \mathcal{D}_c(i, j, v)$ (By the definition of M_c)
$\Leftrightarrow M_c, v \models D_{i,j}\varphi$ (By the definition of $\models$)
Therefore, for any φ, $\varphi \in v$ iff $M_c, v \models \varphi$. □

4.2.1. Corollary. **LD** *is sound and complete in the class of D-models.*

PROOF. In the above theorem, by omitting all of the additional conditions in D-models and the additional theorems in the logical system, we can get the proof. Here we do not want to go to the details. □

We use $\mathbf{LD}^+$ to denote a logic which is obtained from logic **LD** by adding the axiom (DL), i.e., $\neg L_i\bot$. A D-model where all accessibility relations $\mathcal{L}(i)$ are serial is called an LD^+ model. It is also easy to see that:

4.2.2. Corollary. *The logic* $\mathbf{LD}^+$ *is sound and complete in the class of* LD^+*-models*

4.2.3 Decidability and Complexity

In [Halpern&Moses 1992], Halpern and Moses conclude that the complexity of the satisfiability problem for logics of knowledge and beliefs are as follows: The satisfiability problems for $S5$ system with one agent and that for $KD45$ system with one agent are NP-complete; the satisfiability problems for $S5$ system with more than one agent, for $KD45$ system with more than one agent, for K, T, $S4$ systems with one or more than one agent, are PSPACE-complete.

In this section, we focus on the complexity problem about the minimal system **LD**. We leave the complexity problems for other systems for further research. We note that K system is a subsystem of **LD** system. Therefore, we can immediately obtain the lower bound of the complexity of the satisfiability problem by Halpern and Moses's above results. In the following, we just focus on the upper bound of the problem. Considering the facts that D-model is a syntactic approach, and that any formula, even a falsum, can be in any dependence set $\mathcal{D}(i, j, w)$, it is not difficult to obtain the upper bound for the problem by the following trick. We construct a new primitive proposition set by adding countablely new propositions $q_{D_{i,j}\varphi}$ where q is a special symbol and $\varphi \in \mathbf{L_D}$. Namely, let $\Phi^+ = \Phi_0 \cup \{q_{D_{i,j}\varphi} : \varphi \in \mathbf{L_D}\}$. Also

let L^+ be the minimal set which is recursively constructed by the following syntactic rules:(i) $\Phi^+ \subset L_+$; (ii) $\phi, \psi \in L^+, i \in A_n \Rightarrow \neg\phi, \phi \wedge \psi, L_i\phi \in L^+$. We define a function $f : \mathbf{L_D} \to L^+$, which transfers formulas in $\mathbf{L_D}$ into formulas in the logic of knowledge and belief, as follows:

(i) $f(p) = p$, for $p \in \Phi_0$,
(ii) $f(\neg\varphi) = \neg f(\varphi)$,
(iii) $f(\phi \wedge \psi) = f(\phi) \wedge f(\psi)$,
(iv) $f(L_i\varphi) = L_i f(\varphi)$,
(v) $f(D_{i,j}\varphi) = q_{D_{i,j}\varphi}$.

4.2.9. Lemma. *For any formula $\varphi \in \mathbf{L_D}$, φ is satisfiable in a D-model iff $f(\varphi)$ is satisfiable in a K-model (i.e., a Kripke model).*

PROOF. ($\Rightarrow$) Suppose that a formula $\varphi \in \mathbf{L_D}$ is satisfiable in a D-model. Let this D-model $M = \langle W, \mathcal{L}, \mathcal{D}, V\rangle$ and $w \in W$ and $M, w \models \varphi$. We construct a K-model M^+ as follows:

$M^+ = \langle W, \mathcal{L}, V^+\rangle$,
where $V^+(p) = V(p)$, for any $p \in \Phi_0$,
and $V^+(q_{D_{i,j}\varphi}) = \{w \in W : M, w \models D_{i,j}\varphi\}$.

It is easy to see that the above constructed model M^+ indeed is a K-model. Furthermore, we claim that $M, w \models \varphi \Leftrightarrow M^+, w \models f(\varphi)$ for any $\varphi \in \mathbf{L_D}$. We prove it by the induction on the complexity of the formula φ. The cases concerning a primitive proposition $p \in \Phi_0$, and the Boolean connectives $\neg\phi$, $\phi \wedge \psi$, are straightforward.

Case $L_i\varphi$.
$M, w \models L_i\varphi \Leftrightarrow (\forall w' \in W)(\langle w, w'\rangle \in \mathcal{L}(i) \Rightarrow M, w' \models \varphi)$ (Truth condition)
$\Leftrightarrow (\forall w' \in W)(\langle w, w'\rangle \in \mathcal{L}(i) \Rightarrow M^+, w' \models f(\varphi)$ (Induction hypothesis)
$\Leftrightarrow M^+, w \models L_i f(\varphi)$ (Truth condition)
$\Leftrightarrow M^+, w \models f(L_i\varphi)$ (Definition of f)

Case $D_{i,j}\varphi$.
$M, w \models D_{i,j}\varphi \Leftrightarrow w \in V^+(q_{D_{i,j}\varphi})$ (Definition of V^+)
$\Leftrightarrow M^+, w \models q_{D_{i,j}\varphi}$ (Truth condition)
$\Leftrightarrow M^+, w \models f(D_{i,j}\varphi)$ (Definition of f)

($\Leftarrow$) Suppose that $f(\varphi)$ is satisfiable in a K-model $M^+ = \langle W, \mathcal{L}, V^+\rangle$. Thus, there exists a world w such that $M^+, w \models f(\varphi)$. Similarly, we construct a D-model M as follows:

$M = \langle W, \mathcal{L}, \mathcal{D}, V\rangle$

where $\varphi \in \mathcal{D}(i,j,w)$ iff $w \in V^+(q_{D_{i,j}\varphi})$,
$w \in V(p)$ iff $w \in V^+(p)$ for any $p \in \Phi_0$.

Similarly, it is also easy to prove that fact $M, w \models \varphi$ iff $M^+, w \models f(\varphi)$ for any $\varphi \in \mathbf{L_D}$. Here we do not want to go to the details. □

4.2.10. Theorem. *The satisfiability problem of* **LD** *is PSPACE complete*

PROOF. The lower bound comes directly from Halpern and Moses' result. The upper bound follows from the above lemma. Given a formula $\varphi \in \mathbf{L_D}$, we first translate the formula into a formula $f(\varphi)$ in L^+, which can be done within log-time. Then, we use the algorithm for K-satisfiability problem to check the satisfiability of $f(\varphi)$, which can be done within PSPACE by Halpern and Moses's results. If the checking result is "yes", then we say "yes", otherwise we say "no". Therefore, the whole procedure can be done within PSPACE, which concludes our theorem. □

4.2.11. Corollary. *The provability problem of* **LD** *is PSPACE complete.*

PROOF. Immediately from the fact that coPSPACE=PSPACE. □

4.2.12. Corollary. *The problems of the satisfiability and the provability of* **LD** *are decidable.*

4.3 Lij-Model: An Adapted Possible World Approach

4.3.1 Semantics and Lij Logics

The D-models of belief dependence provide a syntactic approach, which to some extent does not coincide with possible world semantics for epistemic logics. Moreover, L-models of belief dependence, which are based on general epistemic logics, suffer from the problem that the dependency operator can not be handled with ease. Therefore, we present a third logic for belief dependence. The ideas is to adapt possible world semantics for modeling belief dependence by directly introducing sub-belief structures. We call all logics based on this approach *Lij logics*. Formally, we have the following definition:

4.3.1. Definition. (Belief dependence Lij-model) *A belief dependence Lij-model is a tuple* $M = \langle W, \mathcal{L}, V \rangle$

where W is a set of possible worlds, V is a truth assignment as usual, and $\mathcal{L}$: $\mathbf{A_n} \times \mathbf{A_n} \to \mathcal{P}(W \times W)$, which consists of $n \times n$ binary accessibility relations on W.

The relation $\models$ is similarly defined inductively as follows:
$M, w \models p$, where p is a primitive proposition, iff $w \in V(p)$,
$M, w \models \neg\varphi$ iff $M, w \not\models \varphi$
$M, w \models \varphi_1 \wedge \varphi_2$ iff $M, w \models \varphi_1$ and $M, w \models \varphi_2$,
$M, w \models L_{i,j}\varphi$ iff $M, w' \models \varphi$ for all w' such $\langle w, w' \rangle \in \mathcal{L}(i, j)$.

For a Lij model $M = \langle W, \mathcal{L}, V \rangle$, a tuple $\langle W, \mathcal{L} \rangle$ is called a *Lij frame*. Satisfiability and validity relation in the Lij approach are defined as usual. $L_{i,j}\varphi$ means that due to agent j, agent i believes the formula φ. In Lij-models, we give $L_{i,i}\varphi$ its general epistemic interpretation, namely, $L_i\varphi$. Just as in the cases of epistemic logic, we generally hope that the axiom $L_i\varphi \to L_iL_i\varphi$ holds. Similarly, for sub-beliefs, we generally hope that the axiom $L_{i,j}\varphi \to L_iL_{i,j}\varphi$ holds. In order to formulate those properties, we need the following definitions:

4.3.2. Definition. (Left-closed accessibility relations) *For any Lij model $M = \langle W, \mathcal{L}, V \rangle$, an accessibility relation $\mathcal{L}(i, j)$ is a left-closed relation, if $\mathcal{L}(i, i) \circ \mathcal{L}(i, j) \subseteq \mathcal{L}(i, j)$ holds.*

4.3.3. Proposition. *For any Lij model in which every accessibility relation is left-closed, the axiom $L_{i,j}\varphi \to L_iL_{i,j}\varphi$ holds.*

4.3.4. Definition. (Almost-Euclidean accessibility relations) *For any Lij-model $M = \langle W, \mathcal{L}, V \rangle$, an accessibility relation $\mathcal{L}(i, j)$ is an almost-Euclidean relation, if $\langle t, u \rangle \in \mathcal{L}(i, j)$ whenever $\langle s, u \rangle \in \mathcal{L}(i, j)$ and $\langle s, t \rangle \in \mathcal{L}(i, i)$.*

4.3.5. Proposition. *For any Lij-model in which every accessibility relation is almost-Euclidean, the axiom $\neg L_{i,j}\varphi \to L_{i,i}\neg L_{i,j}\varphi$ holds.*

4.3.6. Definition. *For any accessibility relation $R \subseteq W \times W$, R is said to be:*
(a) a **D***-relation, if it is serial.*
(b) a **4***-relation, if it is transitive.*
(c) a **5***-relation, if it is Euclidean.*
(d) a **4***-relation, if it is a left-closed relation.*
(e) a **5***-relation, if it is an almost-Euclidean relation.*

4.3.7. Definition. (D4*5* Lij-model) *An Lij-model $M = \langle W, \mathcal{L}, V \rangle$ is a* **D4*5*** *Lij-model, if every accessibility relation on W is serial, left-closed, and almost-Euclidean.*

4.3.8. Theorem. **Lij5**$^-$**+D** *belief dependence logics are sound and complete in the class of* **D4*5*** *Lij-models.*

Proof. Soundness is evident, and completeness can be proved analogous to the proof for the D-models. We define the canonical structure $M_c = \langle S, \mathcal{L}_c, V \rangle$ as follows:

$W = \{v : v \text{ is a maximal consistent set}\}$,
$v \in V(p)$ iff $p \in v$,

$\mathcal{L}_c(i,j) = \{\langle v, w\rangle : L^-_{i,j}(v) \subseteq w\}$.
where $L^-_{i,j}(v) \stackrel{\text{def}}{=} \{\varphi : L_{i,j}\varphi \in v\}$.

First, we show that M_c is a **D4*5*** Lij-model. Axiom (Lij3) guarantees every $\mathcal{L}_c(i,j)$ is serial. For any $\langle v, w\rangle \in \mathcal{L}_c(i,i)$, and $\langle w, w'\rangle \in \mathcal{L}_c(i,j)$,

We have $L^-_{i,i}(v) \subseteq w$ and $L^-_{i,j} \subseteq w'$.
$L_{i,j}\varphi \in v \Rightarrow L_iL_{i,j} \in v$ (Axiom(Lij4) and lemma(2.c))
$\Rightarrow L_{i,j}\varphi \in w$ $(L^-_{i,i}(v) \subseteq w)$
$\Rightarrow \varphi \in w'$ $(L^-_{i,j}(w) \subseteq w')$

Therefore, every accessibility relation is a **4***-relation. Furthermore, for any $\langle v, w\rangle \in \mathcal{L}_c(i,j)$, and $\langle v, w'\rangle \in \mathcal{L}_c(i,i)$,

We have $L^-_{i,j}(v) \subseteq W$ and $L^-_{i,i}(v) \subseteq w'$.
$L_{i,j}\varphi \in w' \Rightarrow L_{i,i}L_{i,j}\varphi \in v$ $(L^-_{i,i}(v) \subseteq w')$
$\Rightarrow \neg L_{i,i}\neg L_{i,j}\varphi \in v$ (Axiom $L_{i,i}\psi \rightarrow \neg L_{i,i}\neg\psi$)
$\Rightarrow L_{i,j}\varphi \in v$ (Axiom $(Lij5)$)
$\Rightarrow \varphi \in w$ $(L^-_{i,j}(v) \subseteq w)$

Therefore, every accessibility relation is a **5***-relation, i.e., M_c is a **D4*5*** Lij-model. Similarly to the proof in the D-model approach, we can show with induction that $M_c, v \models \varphi$ iff $\varphi \in v$. The only non-trivial step is for $L_{i,j}\varphi$. Similar to that in D-model, we also can prove the corresponding lemma which says that, if a maximal consistent set w contains a formula $\neg L_{i,j}\varphi$, then $L^-_{i,j}(w) \cup \{\neg\varphi\}$ is consistent. Therefore, the induction proof can go through. Details are routine. □

4.3.1. Corollary. *Logic* **LIJ** *is sound and complete in the class of Lij-models.*

PROOF. By omitting the additional conditions in the semantic models and the additional axioms in the logic system, similar to the proof in the above theorem, we can have the proof. □

4.3.2 Decidability and Complexity

The satisfiability problems of Lij logics are stated as follows: Given a formula φ, is φ satisfiable in some world of some Lij model. We note that general epistemic logic is a subsystem of Lij logic because we interpret $L_{i,i}\varphi$ as $L_i\varphi$. Therefore, we can use the result about the lower bound of the complexity problem of standard knowledge logics for Lij logics. Furthermore, note that for the single agent case, i.e., $|A_n| = 1$, all Lij logics collapse into the standard knowledge logics. Therefore, the complexity results for one agent in the standard knowledge logics are the results of the corresponding

Lij logics. In the following, we focus on the problem for **LIJ** logic with more than one agent. According to Halpern and Moses's work in [Halpern&Moses 1992], the complexity of the satisfiability problem of **K** system is PSPACE-complete. Therefore, we immediately get the result that the satisfiability problem of **LIJ** logic is PSPACE-hard, which provides a lower bound for the problem.

In order to achieve the upper bound of the problem, we would like to use the tableau method for **LIJ** logic, which is a generalization of Halpern and Moses' work in [Halpern&Moses 1992].

A *propositional tableau* is a set T of formulas such that
(1) if $\neg\neg\psi \in T$, then $\psi \in T$,
(2) if $\psi \wedge \psi' \in T$, then both $\psi, \psi' \in T$,
(3) if $\neg(\psi \wedge \psi') \in T$, then either $\neg\psi \in T$ or $\neg\psi' \in T$, and
(4) it is not the case that both ψ and $\neg\psi$ are in T for some formula ψ.

We say that T is a *propositional tableau for* φ if T is a propositional tableau and $\varphi \in T$. It is easy to see that the propositional formula φ is satisfiable if and only if there is a propositional tableau for φ.

We now extend the notion of a propositional tableau to a tableau for **LIJ** logic. An *Lij-tableau* is a tuple $T = \langle W, \mathcal{L}, L\rangle$, where, $\langle W, \mathcal{L}\rangle$ is a Lij frame, while L is *labeling function* that associates with each world $w \in W$ a set $L(w)$ of formulas such that

(1) $L(w)$ is a propositional tableau,
(2) $L_{i,j}\psi \in L(w)$ and $\langle w, w'\rangle \in \mathcal{L}(i,j) \Rightarrow \psi \in L(w')$,
(3) $\neg L_{i,j}\psi \in L(w) \Rightarrow \exists w'(\langle w, w'\rangle \in \mathcal{L}(i,j)$ and $\neg\psi \in L(w')$.

We say that $T = \langle W, \mathcal{L}, L\rangle$ is an *Lij tableau for* φ if T is an Lij tableau and $\varphi \in L(w)$ for some world $w \in W$. A formula set T is said to be a *fully expanded formula set* if for any $\varphi \in T$ and subformula ψ of φ, either $\psi \in T$ or $\neg\psi \in T$. An Lij tableau $T = \langle W, \mathcal{L}, L\rangle$ is said to be a *fully-expanded Lij tableau for* φ if for some world $w \in W$, $\varphi \in L(w)$ and $L(w)$ is fully expanded formula set.

4.3.9. Proposition. *The formula φ is Lij satisfiable iff there is a fully expanded Lij tableau for φ.*

PROOF. If φ is Lij satisfiable, suppose it is satisfied in the model $M = \langle W, \mathcal{L}, V\rangle$. Let $T = \langle W, \mathcal{L}, L\rangle$, where $L(w) = \{\psi : M, w \models \psi\}$. It is easy to see that T is a fully expanded Lij tableau for φ. For the converse, suppose that $T = \langle W, \mathcal{L}, L\rangle$ is a fully expanded Lij tableau for φ. Then, we know that there exists a world $w \in W$ such that $\varphi \in L(w)$. We construct a model $M = \langle W, \mathcal{L}, V\rangle$ where $w' \in V(p)$ iff $p \in L(w')$ for each primitive proposition p and each world $w' \in W$.

We now show by induction on the structure of formulas that if $\psi \in Sub(\varphi)$, then $M, w \models \psi \Leftrightarrow \psi \in L(w)$, where $Sub(\varphi)$ is the subformula set of φ. The proof for the cases ψ is a primitive proposition p_i, a negation of a formula, $\neg\psi'$, and a conjunction $\psi_1 \wedge \psi_2$, are straightforward. The non-trivial steps are:

Case $L_{i,j}\phi$,

($\Rightarrow$)
$M, w \models L_{i,j}\phi$ and $L_{i,j}\phi \notin L(w)$
$\Rightarrow \forall w'(\langle w, w'\rangle \in \mathcal{L}(i,j) \Rightarrow M, w' \models \phi)$ and $L_{i,j}\phi \notin L(w)$ (Truth condition)
$\Rightarrow \forall w'(\langle w, w'\rangle \in \mathcal{L}(i,j) \Rightarrow \phi \in L(w'))$ and $L_{i,j}\phi \notin L(w)$ (Induction hypothesis)
$\Rightarrow \forall w'(\langle w, w'\rangle \in \mathcal{L}(i,j) \Rightarrow \phi \in L(w'))$ and $\neg L_{i,j}\phi \in L(w)$ ($\varphi \in L(w), L_{i,j}\psi \in Sub(\varphi)$, and fully expand condition)
$\Rightarrow \forall w'(\langle w, w'\rangle \in \mathcal{L}(i,j) \Rightarrow \phi \in L(w'))$ and $\exists w'(\langle w, w'\rangle \in \mathcal{L}(i,j)$ and $\neg\phi \in L(w'))$ ((3) of Lij tableau definition)
$\Rightarrow$ **False** ((3) of propositional tableau definition)

($\Leftarrow$)
$L_{i,j}\phi \in L(w)$ and $M, w \not\models L_{i,j}\phi$
$\Rightarrow L_{i,j}\phi \in L(w)$ and $\exists w'(\langle w, w'\rangle \in \mathcal{L}(i,j)$ and $M, w' \models \neg\phi)$ (Truth condition)
$\Rightarrow L_{i,j}\phi \in L(w)$ and $\exists w'(\langle w, w'\rangle \in \mathcal{L}(i,j)$ and $\neg\phi \in L(w'))$ (Induction hypothesis)
$\Rightarrow \phi \in L(w')$ and $\neg\phi \in L(w')$ ((3) of Lij tableau definition)
$\Rightarrow False$ ((4) of tableau definition)

Therefore, we conclude that if $\psi \in Sub(\varphi)$, then $M, w \models \psi \Leftrightarrow \psi \in L(w)$, which implies that there is a fully expanded Lij tableau for φ, then φ is Lij satisfiable. □

Given a formula φ, we now present an algorithm that attempts to construct a fully expanded Lij tableau for φ. We show that the construction succeeds if and only if φ is Lij satisfiable. Finally, we show that there is an algorithm that checks whether our tableau construction succeeds that runs in space polynomial in $|\varphi|$.

The constructed tableau is actually a tree with some labeled edges. The tableau construction consists of four independent procedures. The first procedure expands a set of formulas to a propositional tableau. The second constructs a fully expanded propositional tableau for φ. The third takes a node whose label is a fully expanded propositional tableau and creates successors to the node so as to satisfy clause (3) of the definition of Lij tableau. The fourth procedure checks for satisfiable labels.

If a set T of formulas is not a propositional tableau, then ψ is a *witness* to this if $\psi \in T$ and one of clauses (1)-(3) in the definition of propositional tableau does not apply to ψ. We assume that the formulas are ordered in some way so that it makes sense to choose the "least witness" if there is a witness. We say that a set T is *blatantly inconsistent* if, for some formula ψ, both ψ and $\neg\psi$ are in T. We say that a node w is a *leaf* iff w has no successors.

The fully expanded Lij tableau construction for φ_0:

Step 1. Construct a tree consisting of a single node w_0 (the "root"), with $L(w_0) = \{\varphi_0\}$.
Step 2. Repeat until none of (2.1)-(2.4) below applies:
(2.1) *Forming a propositional tableau*: If w is a leaf of the tree, $L(w)$ is not blatantly inconsistent, $L(w)$ is not a propositional tableau, and ψ is the least witness to this fact, then:
(2.1.1) if ψ is of the form $\neg\neg\psi'$, then creates a successor w' of w (i.e., add a node w' to the tree and an edge from w to w') and set $L(w') = L(w) \cup \{\psi'\}$,
(2.1.2) if ψ is of the form $\psi_1 \wedge \psi_2$, then create a successors w' of w and set $L(w') = L(w) \cup \{\psi_1, \psi_2\}$,
(2.1.3) if ψ is of the form $\neg(\psi_1 \wedge \psi_2)$, then create two successors w_1 and w_2 of w and set $L(w_1) = L(w) \cup \{\neg\psi_i\}, i = 1, 2$.
(2.2) *Forming a fully expanded propositional tableau*: If w is a leaf of the tree, $L(w)$ is not blatantly inconsistent, $L(w)$ is not a fully expanded propositional tableau, and ψ is the least witness to this fact, then create two successors w' and w" of w and set $L(w') = L(w) \cup \{\psi\}$ and $L(w") = L(w) \cup \{\neg\psi\}$.
(2.3) *Creating successor nodes*: If w is a leaf of the tree, $L(w)$ is not blatantly inconsistent, and $L(w)$ is a fully expanded propositional tableau, then for each formula of the form $\neg L_{i,j}\psi \in L(w)$, create a (i,j)-successor node w' (i.e., add the node w' to the tree and an edge from w to w' labeled (i, j)) and let $L(w') = \{\neg\psi\} \cup \{\rho : L_{i,j}\rho \in L(w)\}$;
(2.4) *Making nodes 'satisfiable'*: If w is not marked 'satisfiable' then mark w 'satisfiable' if either
(2.4.1) $L(w)$ is not a fully expanded propositional tableau and w' is marked 'satisfiable' for some successor w' of w,
(2.4.2) $L(w)$ is a fully expanded propositional tableau, there are no formulas of the forms $\neg L_{i,j}\psi \in L(w)$, and $L(w)$ is not blatantly inconsistent, or
(2.4.3) $L(w)$ is a fully expanded propositional tableau, w has successors and all of them are marked 'satisfiable'.
Step 3. If the root of the tree is marked 'satisfiable', then return "φ_0 is satisfiable"; otherwise return "φ_0 is unsatisfiable".

4.3.10. Lemma. *For all formula φ, the Lij tableau construction terminates.*

PROOF. It is straightforward from the above construction. Suppose that $|\varphi| = m$. Note that the above construction guarantees that for any node w in the tree, $L(w)$ consists only of formulas in $Sub^+(\varphi)$ where $Sub^+(\varphi)$ denotes the subformula and its simple negation set of φ. So, $|L(w)| \leq 2m$. Furthermore, note that with the tree process from root to leaf, the complexity of labeling set becomes more simpler. □

We call a node w for which $L(w)$ is not a fully expanded propositional tableau *an internal node*; otherwise we call w a *world*. From the construction, we know that for

any world w, if there exists a successor w' of w, then w' can arrive at another world w" via a chain of internal nodes; furthermore, the chain of internal nodes is empty if w' already is a world. We call w" is a real successor of w if w' is a successor of w.

4.3.11. THEOREM. *A formula φ is Lij satisfiable iff the Lij tableau construction for φ returns "φ is satisfiable".*

PROOF. ($\Leftarrow$) Suppose that the Lij tableau construction for φ returns "φ is satisfiable ". We construct a fully expanded Lij tableau $T = \langle W, \mathcal{L}, L\rangle$ for φ as follows:
W consists of all of the worlds w in the construction which are marked "satisfiable"; $\langle w, w'\rangle \in \mathcal{L}(i, j)$ iff w' is a real (i,j)-successor of w. It is easy to see that T is a fully expanded Lij tableau for φ. According to the proposition, we know that φ is Lij satisfiable.

($\Rightarrow$) For a node w, let ψ_w be the conjunction of all the formulas in $L(w)$. We prove that if a node w in the construction is not marked "satisfiable", then ψ_w is Lij unsatisfiable.

We show that by induction on the height of w (i.e., then length of the longest path from w to a leaf.) If w has no successors, then w is not marked "satisfiable" if and only if $L(w)$ is blatantly inconsistent from the step (2.4.2). Therefore, in this case, ψ_w is Lij unsatisfiable. If w has successors and $L(w)$ is not a fully expanded propositional tableau, from step (2.4.1), it follows that w is not marked "satisfiable" if and only if none of w's successors is marked "satisfiable". By the induction hypothesis, it follows that $\psi_{w'}$ is Lij unsatisfiable for every successor w' of w. It is easy to see that ψ_w is Lij unsatisfiable. If w has successors and $L(w)$ is a fully expanded propositional tableau, from step (2.4.3) it follows that w is not marked "satisfiable" if and only if some of successors w' of w is not marked "satisfiable. By construction, there exists a formula $\neg L_{i,j}\psi \in L(w)$ such that $L(w') = \{\rho : L_{i,j}\rho \in L(w)\} \cup \{\neg\psi\}$. By the induction hypothesis, it follows that $\psi_{w'}$ is Lij unsatisfiable. So, there exist formulas $L_{i,j}\rho_0, L_{i,j}\rho_1, ..., L_{i,j}\rho_k \in L(w)$ such that $\rho_0 \wedge \rho_1 \wedge ... \wedge \rho_k \wedge \neg\psi$ is Lij unsatisfiable.

However, $\rho_0 \wedge \rho_1 \wedge ... \wedge \rho_k \wedge \neg\psi$ is Lij unsatisfiable
$\Rightarrow \neg\rho_0 \vee \neg\rho_1 \vee ...\neg\rho_k \vee \psi$ is valid
$\Rightarrow ((\rho_1 \wedge ...\rho_k) \rightarrow \psi)$ is valid
$\Rightarrow L_{i,j}\psi \in L(w)$
$\Rightarrow \psi_w$ is Lij unsatisfiable (Since $\neg L_{i,j}\psi \in L(w)$). □

4.3.12. THEOREM. *There is an algorithm for deciding satisfiability of LIJ logic formulas that runs in polynomial space.*

PROOF. From the construction, it is easy to see that we only need polynomial space to implement the algorithm. □

Approaches	*Efficiency*	*Intuition*	*Avoidance of LO*
D-model	Yes	No	Yes
L-model	No	Yes	No
Lij-model	Yes	Yes	No

Where LO means the problem of logical omniscience.

Figure 4.2: Summaries about Approaches

4.3.13. THEOREM. *The satisfiability problem for* **LIJ** *logic is PSPACE-complete.*

PROOF. The lower bound of the complexity is implied by Halpern and Moses work in [Halpern&Moses 1992]. The upper bound of the complexity is obtained from the above theorem. □

4.3.2. COROLLARY. *The provability problem for* **LIJ** *logic is PSPACE-complete.*

PROOF. Straightforward from the fact coPSPACE=PSPACE. □

4.3.3. COROLLARY. *The problem of satisfiability and the problems of validity of* **LIJ** *logic are decidable.*

4.4 A Brief Comparison

So far several semantic approaches for belief dependency logic have been proposed. All of these approaches capture certain properties of belief dependence. There are many different criteria for appraising these approaches. I suggest the following criteria:

i) *Adequacy of Efficiency*: The approach can efficiently formalize fundamental features such as closure, suspicion, indirect dependence, and role-appraisal.
ii) *Adequacy of Intuition*: The approach is intuitively appealing.
iii) *Avoidance of Logical Omniscience*: Approaches do not suffer from the problem of logical omniscience.

The D-model is based on a syntactic strategy amalgamated with a possible world approach. Therefore, it can avoid the problem of logical omniscience. The L-model represents a general epistemic logic approach, which fails, however, to capture some important features of the dependency operator. The Lij-model seems to be more reasonable and acceptable, since it can capture many intuitive properties concerning the dependency operator, although the approach suffers from the problem of logical omniscience, just as the L-model. The comparison is shown in the figure 4.2.

From this comparison, we know that each semantic model has its own advantages and disvantages. Deciding which semantics approach should be used depends on the applications.

Chapter 5

Belief Dependence, Revision, and Persistence

5.1 Belief Maintenance

Starting this chapter, I consider the second phase of the formalism. I will study the problem of belief dynamics for a rational agent which has some sub-beliefs from someone else. However, as mentioned before, contemporary cognitive psychology has not yet offered a theory to specify the process how and when an rational agent transfers her sub-beliefs into her incorporated beliefs. One of the possibilities to formalize the process is to borrow some technique in philosophy. The belief dynamics theories proposed by Gärdenfors and Makinson et al.[Gärdenfors 1988, Gärdenfors 1990, Gärdenfors&Makinson 1988] give a lot of insights on the problem. I will use some of their ideas.

There are several different strategies for an rational agent's belief dynamics . When an rational agent meets some sub-belief, she may simply accept the sub-belief as her own belief. If the new belief is consistent with her original belief, she just expands the original one by simply accepting the new one. This process is called *belief expansion*. A more complicated case is that the new belief is inconsistent with her original belief. Under that situation, she has to remove some of her original beliefs to make the new one consistent. We call the process *belief revision*. Sometimes the rational agent may like to keep her original belief unchanged, although she meets some new sub-belief, the process is called *belief persistence*. Even in some case, the rational agent may even remove some part of her original belief without accepting any new one. We call the process *belief contraction*. The processes mentioned above constitute the options of belief dynamics. Therefore, we need a general notion of belief maintenance. We use the symbol $\triangle$ to denote belief maintenance operators. Therefore, one of tasks in this chapter is to offer a general formalism for the belief maintenance operation under the framework of the belief dependence.

Belief revision has been one of major problems for belief management techniques and knowledge representation systems [Gärdenfors 1988, Gärdenfors&Makinson 1988,

Martin&Shapiro 1986, Nebel 1990]. Belief revision and belief persistence are very important features for flexible and intelligent knowledge based systems. We are looking for an intuitive approach by which the choice between belief revision and belief persistence can be determined by rational agents. In other words, we are looking for a formalism where one can mechanically compute, given a description of the state of the world whether an agent will revise her beliefs or persist when confronted with new incompatible information.

In this chapter, a rational formalism which makes the choice between belief revision and belief persistence computable is presented. Our approach heavily depends on the theory of belief revision. Therefore, in the following, we first present a brief overview the theory of belief revision.

5.2 A Bit of Belief Revision Theories

There exist three kinds of update operations used for belief maintenance [Gärdenfors 1988, Gärdenfors&Makinson 1988]. They are: *Expansion, Revision*, and *Contraction*. Formally,

Expansion: A new proposition φ is added to a given knowledge set K. Formally, the knowledge set that results from expanding K by a proposition φ is denoted $K + \varphi$.

Revision: A new proposition which is inconsistent with a knowledge set K is added, but in order to keep the resulting set consistent some of the old propositions in K must be removed. The resulting of revising K by a proposition φ is denoted $K\dot{+}\varphi$.

Contraction: A proposition in K is retracted without adding any new proposition. The resulting knowledge set of contracting K with respect to the formula φ is denoted $K\dot{-}\varphi$.

Suppose the added proposition is φ, the knowledge set K, and the base of the knowledge set be B. A possible definition for the expansion operation reads:

$$Cn(B) + \varphi \stackrel{\text{def}}{\Longleftrightarrow} Cn(Cn(B) \cup \{\varphi\}) = Cn(B \cup \{\varphi\}).$$

where Cn is the *consequence operation*, a closure operator which maps sets of propositions to set of propositions and which has the following properties:

inclusion $A \subseteq Cn(A)$.
idempotency $Cn(A) = Cn(Cn(A))$.
monotonicity $Cn(A) \subseteq Cn(B)$, whenever $A \subseteq B$.

For revision and contraction, the following condition, called *Levi identity*, is generally required to be satisfied:

$$K\dot{+}\varphi = Cn((K\dot{-}\neg\varphi) \cup \{\varphi\}).$$

Although the revision and contraction operations to a large extent are a matter of pragmatics, and seem to be beyond the scope of logical analysis, there have been

proposed a number of intuitive plausible constraints in the work of Alchourrón, Gärdenfors, and Makinson; see [Gärdenfors 1988, Gärdenfors&Makinson 1988]. Gärdenfors presents a set of constraints on the contraction operator which are called the *Gärdenfors Postulates*:

(1) [Closure] $A\dot{-}\varphi$ is a closed theory.

(2) [Inclusion] $A\dot{-}\varphi \subseteq A$.

(3) [Vacuity] If $\varphi \notin A$, then $A\dot{-}\varphi = A$.

(4) [Success] If $\varphi \notin Cn(\emptyset)$, then $\varphi \notin A\dot{-}\varphi$.

(5) [Extensionality] If $Cn(\varphi) = Cn(\psi)$, then $A\dot{-}\varphi = A\dot{-}\psi$.

(6) [Recovery] $A \subseteq Cn((A\dot{-}\varphi) \cup \{\varphi\})$.

(7) [Intersection] $(A\dot{-}\varphi) \cap (A\dot{-}\psi) \subseteq A\dot{-}(\varphi \wedge \psi)$.

(8) [Conjunction] If $\varphi \notin A\dot{-}(\varphi \wedge \psi)$, then $A\dot{-}(\varphi \wedge \psi) \subseteq A\dot{-}\varphi$.

The correponding set of constraints on the revision operator is:

(a) [Closure] $A\dot{+}\varphi$ is a closed theory.

(b) [Inclusion] $A\dot{+}\varphi \subseteq A + \varphi$.

(c) [Vacuity] If $\neg\varphi \notin A$, then $A + \varphi \subseteq A\dot{+}\varphi$.

(d) [Success] $\varphi \in A\dot{+}\varphi$.

(e) [Consistency] If $\bot \in A\dot{+}\varphi$, then $\neg\varphi \in Cn(\emptyset)$.

(f) [Extensionality] If $Cn(A) = Cn(B)$, then $A\dot{+}\varphi = B\dot{+}\varphi$.

(g) [Conjunctive Inclusion] $A\dot{+}(\varphi_1 \wedge \varphi_2) \subseteq ((A\dot{+}\varphi_1) + \varphi_2)$.

(h) [Conjunctive Vacuity] If $\neg\varphi_2 \notin A\dot{+}\varphi_1$, then $((A\dot{+}\varphi_1) + \varphi_2) \subseteq A\dot{+}(\varphi_1 \wedge \varphi_2)$.

Moreover, one may define the contraction operation in terms of the revision operation. The following identity is called *Harper identity*:

$$A\dot{-}\varphi = (A\dot{+}\neg\varphi) \cap A.$$

The idea here is that one may retract φ from A by first revising to include the negation of φ, and then intersecting with the original theory A. By the Levi identity and the Harper identity, Gárdenfors and Makinson show that the two sets of postulates

support each other by the following properties: If a revision function $\dot{+}$ is defined by the Levi identity from a contraction function satisfying the postulates (1) to (4) and (6) to (8), then the function $\dot{+}$ satisfies (a) to (h).
If a contraction function $\dot{-}$ is defined by the Harper identity from a revision function satisfying the postulateds (a) to (h), then the function $\dot{-}$ satisfies (1) to (8).

5.3 Belief Maintenance Under the Framework of Logics of Belief Dependence

As argued in chapter 2, we believe that an appropriate procedure for formalizing information assimilation should pass two phases: compartmentalization and incorporation of information. Recall that compartmentalized information consists of the fragments of information which are accepted and remembered as isolated beliefs but which are treated differently from those beliefs which are completely believed. Whereas incorporated information consists of those beliefs which are completely believed by the agents. In the logic for belief dependence, compartmentalized information corresponds to sub-beliefs $L_{i,j}\varphi$ for agent i, Whereas incorporated information corresponds to standard beliefs $L_i\varphi$ of agent i.

More formally, the stages can be described as follows: if an agent i relies on another agent j about believing formula φ, and agent j believes φ, then agent i will accept the belief φ. However, in the first stage of assimilation of knowledge and beliefs, agent i only accepts φ as a sub-belief $L_{i,j}\varphi$. Agent i does not necessarily accept the belief as her own incorporated belief.

In the second stage of assimilation of knowledge and belief agent i is supposed to change her compartmentalized belief into incorporated belief, i.e., change $L_{i,j}\varphi$ into $L_i\varphi$. However, if we just simply transform the sub-belief, then we will find that the resulting beliefs may turn out to be inconsistent. In order to avoid such conflicts, some old beliefs in $L_i\varphi$ must be removed. Alternatively, the agent might reject the new information in order to avoid the inconsistency. Thus, under those circumstances, we must use the belief update operations revision $\dot{+}$ and contraction $\dot{-}$ to describe the process precisely, and we will need further information in order to decide which operator will be invoked under which circumstances.

For a further formalization we need some further notations. We call a consistent formula set a *belief set*. If K is a formula set, we define that $L^{-}_{i,j}(K) \stackrel{\text{def}}{=} \{\psi : K \models L_{i,j}\psi\}$, denoting the set of agent i's sub-belief indexed by j, and $L^{-}_{i}(K) \stackrel{\text{def}}{=} \{\varphi : K \models L_i\varphi\}$, denotes agent i's belief set. The belief maintenance operator is denoted by $\triangle$.

Let $\mathbf{K}$ is the set of all belief sets. A belief maintenance operation $\triangle : \mathbf{K} \times \mathbf{L} \to \mathbf{K}$ is a function assigning a belief set $\triangle(K, \varphi)$ to any belief set $K \in \mathbf{K}$ and each formula φ in $\mathbf{L}$. We shall write $K\triangle\varphi$ as an alternative representation for $\triangle(K, \varphi)$. Intuitively, $K\triangle\varphi$ denote the agent's resultant belief set after she faces the new information φ with respect to her original belief set K.

5.4 Types of Belief Maintenance Operation

The function $\triangle$ can be defined arbitrarily. However, frequently we are interested in some special form of the belief maintenance operator $\triangle$. Especially, we are interested in the form of the function $\triangle$ in which the rational agent checks whether or not some special formula φ_i is implied by her original belief set X when she faces the new information ρ_i'. If that holds, then the result of $\triangle(X, \rho_i')$ is a new belief set Y_i. Therefore, this function is of the following form, which is called *type 1 belief maintenance operation* .

$$\triangle(X,\rho) = \begin{cases} Y_{i_1} & \text{if } X \models \varphi_{i_1}, \rho = \rho'_{i_1} \\ Y_{i_2} & \text{if } X \models \varphi_{i_2}, \rho = \rho'_{i_2} \\ \ldots & \ldots \\ Y_{i_{n_1}} & \text{if } X \models \varphi_{i_{n_1}}, \rho = \rho'_{i_{n_1}} \\ Y_{i_{n_1+1}} & \text{if } X \not\models \varphi_{i_{n_1+1}}, \rho = \rho'_{i_{n_1+1}} \\ \ldots & \ldots \\ Y_{i_{n_2}} & \text{if } X \not\models \varphi_{i_{n_2}}, \rho = \rho'_{i_{n_2}} \\ X & \textbf{otherwise} \end{cases}$$

Actually, we are more interested in a more restricted form of the the function $\triangle$ where the rational agent only cares to check for formulas φ_i which may belong to her original belief set X when she faces the new information ρ_i' but does not check for beliefs which she does not support. This form of the function is called *type 2 belief maintenance operation* .

$$\triangle(X,\rho) = \begin{cases} Y_{i_1} & \text{if } X \models \varphi_{i_1}, \rho = \rho'_{i_1} \\ Y_{i_2} & \text{if } X \models \varphi_{i_2}, \rho = \rho'_{i_2} \\ \ldots & \ldots \\ Y_{i_n} & \text{if } X \models \varphi_{i_n}, \rho = \rho'_{i_n} \\ X & \textbf{otherwise} \end{cases}$$

Furthermore, in many applications, we are interested in an even more restricted form of the belief maintenance operator $\triangle$. where the rational agent checks whether or not some special formula φ_i belongs to her original belief set X when she faces the new information ρ_i', rather than testing for support by her beliefs. If that holds, then the result of $\triangle(X, \rho_i')$ is a new belief set Y_i. A function of the form below is called *type 3 belief maintenance operation* .

$$\triangle(X,\rho) = \begin{cases} Y_{i_1} & \text{if } \varphi_{i_1} \in X, \rho = \rho'_{i_1} \\ Y_{i_2} & \text{if } \varphi_{i_2} \in X, \rho = \rho'_{i_2} \\ \ldots & \ldots \\ Y_{i_n} & \text{if } \varphi_{i_n} \in X, \rho = \rho'_{i_n} \\ X & \textbf{otherwise} \end{cases}$$

Let **TYPE1**, **TYPE2**, and **TYPE3** be the set of belief maintenance operations which can be defined by type 1, type 2, and type 3 respectively. It is easy to see that the following inclusions hold.

TYPE3⊂TYPE2⊂TYPE1

For the type 3 belief maintenance operation, we can represent the rule by a formula of simplified format:

$$\varphi_{i_j} \Rightarrow X \triangle \rho'_{i_j} = Y_{i_j}.$$

Each rule represents some case in the definition of the function. We omit the rule representation concerning the "otherwise" case because it is a default. Intuitively, each rule with above form says that if φ_i holds in the belief set X, then the result of the maintenance with the new information ρ'_i is Y_i.

Next we would like to use the ordinary update operations such as revision, contraction, and expansion $+$, to define the belief maintenance operation. A complication is that belief revision and contraction functions are not unique, and therefore we must select one of the possiple revision functions which satisfies some of the Gärdenfors postulates as our revision function; we do not care about the details of the belief revision and contraction operation at this moment. Although it is not necessary to require that the selected revision function meets all of the postulates, we require that it at least meets the success constraint (d). Let the selected revision function be $\dot{+}$. Subsequently we define the contraction function $\dot{-}$ by the Harper Identity in terms of the revision function $\dot{+}$. Suppose that $\theta \in \{\dot{+}, \dot{-}, +\}$. Eventually this leads to a definition of the operation in question of the following *type 4 form*:

$$\triangle(X, \rho) = \begin{cases} X\theta\psi_{i_1} & \text{if } \varphi_{i_1} \in X, \rho = \rho'_{i_1} \\ X\theta\psi_{i_2} & \text{if } \varphi_{i_2} \in X, \rho = \rho'_{i_2} \\ \ldots & \ldots \\ X\theta\psi_{i_n} & \text{if } \varphi_{i_n} \in X, \rho = \rho'_{i_n} \\ X & \textbf{otherwise} \end{cases}$$

Let **TYPE4** be the set of the belief maintenance operations which can be represented by type 4 form. It is also easy to see that

TYPE4⊂TYPE3.

We are working with belief maintenance operations for a multi-agent system. Different rational agents may have different belief maintenance strategies. Therefore, we have to state which function corresponds which agent. In many applications, we are only interested in the belief maintenance operation for a special receiver, say, a. Her belief set X normally is $L_a^-(K)$, where K is a knowledge base. Therefore, generally, the definition of the operation in question obtains the following *type 5 form*:

$$\triangle(L_a^-(K),\rho) = \begin{cases} L_a^-(K)\theta\psi_{i_1} & \text{if } \varphi_{i_1} \in L_a^-(K), \rho = \rho'_{i_1} \\ L_a^-(K)\theta\psi_{i_2} & \text{if } \varphi_{i_2} \in L_a^-(K), \rho = \rho'_{i_2} \\ \ldots & \ldots \\ L_a^-(K)\theta\psi_{i_n} & \text{if } \varphi_{i_n} \in L_a^-(K), \rho = \rho'_{i_n} \\ K & \textbf{otherwise} \end{cases}$$

5.5 A Belief Maintenance Strategy Using Logics of Belief Dependence

5.5.1 Update Strategies

In order to deal with different kinds of belief conflicts, we will use the following five types of update strategies[1]:

(Positive-revision)	$L_i^-(K)\triangle L_{i,j}\varphi = L_i^-(K)\dot{+}\varphi$
(Negative-revision)	$L_i^-(K)\triangle L_{i,j}\varphi = L_i^-(K)\dot{+}\neg\varphi$
(Persistence)	$L_i^-(K)\triangle L_{i,j}\varphi = L_i^-(K)$
(Positive-contraction)	$L_i^-(K)\triangle L_{i,j}\varphi = L_i^-(K)\dot{-}\varphi$
(Negative-contraction)	$L_i^-(K)\triangle L_{i,j}\varphi = L_i^-(K)\dot{-}\neg\varphi$

The update strategies of negative-revision and positive-contraction seem to be rather counterintuitive at first sight. However, they describe the behavior of an agent i which doesn't trust his informant j at all; if j believes something this is taken to be a good reason for not accepting it as an incorporated belief, so agent i will rather retract it from his own belief or even add its negation. However, in the sequel of this chapter these two paranoid operators will be disregarded. This leaves, for the belief conflict mentioned above, where $\varphi \in L_{i,j}^-(K)$ and $\neg\varphi \in L_i^-(K)$, we have three reasonable and plausible choices:

1. $\varphi \in L_{i,j}^-(K), \neg\varphi \in L_i^-(K) \Rightarrow L_i^-(K)\triangle L_{i,j}\varphi = L_i^-(K)\dot{+}\varphi$.
   (Agent i accepts the new belief φ on behalf of agent j's believing φ, although agent i originally believes that φ is false.)
2. $\varphi \in L_{i,j}^-(K), \neg\varphi \in L_i^-(K) \Rightarrow L_i^-(K)\triangle L_{i,j}\varphi = L_i^-(K)\dot{-}\neg\varphi$.
   (Agent i retracts her belief that φ is false on behalf of agent j's believing φ.)

3. $\varphi \in L_{i,j}^-(K), \neg\varphi \in L_i^-(K) \Rightarrow L_i^-(K)\triangle L_{i,j}\varphi = L_i^-(K)$.
   (Notwithstanding agent j's belief in φ, agent i persists his belief that φ is false.)

Note that for this special case the alternative of negative-revision, $L_i^-(K)\triangle L_{i,j}\varphi = L_i^-(K)\dot{+}\neg\varphi$, is useless, because $\neg\varphi \in L_i^-(K)$ implies $L_i^-(K)\dot{+}\neg\varphi = L_i^-(K)$. In other words, the strategy coincides with persistence. The same holds for the strategy of positive contraction.

[1] In the following, we make no distinction between $\dot{+}$ and $+$.

The above analysis shows that there exist three different plausible choices for the belief conflict situation under consideration. In order to disambiguate this conflict situation we refine the notion of sub-belief and we introduce three new credibility operators in the formalism: *High-credibility sub-belief, Neutral-credibility sub-belief* and *Low-credibility sub-belief*, denoted $HL_{i,j}\varphi$, $NL_{i,j}\varphi$ and $LL_{i,j}\varphi$ respectively.

The formula $HL_{i,j}\varphi$ means that agent i views agent j as an agent with high credibility on φ, and $NL_{i,j}\varphi$ and $LL_{i,j}\varphi$ mean respectively that with neutral credibility and that with low credibility.

Based on those credibility operators, the three possible outcomes of the above belief conflict are determined as follows:

SH-Pr $\varphi \in L^-_{i,j}(K) \wedge \neg\varphi \in L^-_i(K) \wedge HL_{i,j}\varphi \Rightarrow L^-_i(K) \triangle L_{i,j}\varphi = L^-_i(K) \dot{+} \varphi$.
(Because agent j believes φ and agent j is viewed as an agent with high credibility about φ, agent i accepts the new belief φ, although he originally believes that φ is false.)

SN-Nc $\varphi \in L^-_{i,j}(K), \neg\varphi \in L^-_i(K) \wedge NL_{i,j}\varphi \Rightarrow L^-_i(K) \triangle L_{i,j}\varphi = L^-_i(K) \dot{-} \neg\varphi$.
(Because agent j believes φ and agent j is viewed as an agent with neutral credibility about φ, agent i retracts his original belief that φ is false)

SL-Pe $\varphi \in L^-_{i,j}(K), \neg\varphi \in L^-_i(K) \wedge LL_{i,j}\varphi \Rightarrow L^-_i(K) \triangle L_{i,j}\varphi = L^-_i(K)$.
(Although agent j believes φ while agent i believes that φ is false, agent i persists in his belief, because of agent j's low credibility.)

5.5.2 Role Analysis

In the previous section we have solved the problem of determining which of the three meaningful belief revision strategies will be invoked in the case of a conflict between the new information and the agent's previous belief by refining the notion of sub-belief into three new notions. This means, however that we now must cope with a new problem involving these credibility operators; we must invent a strategy which decides which credibility operators will be the result of the first stage of the process of information assimilation under which circumstances. This choice must be based on the logic of belief dependence and not invoke other information outside the formalism.

Below are some plausible axioms for credibility operators. The credibility operator $HL_{i,j}\varphi$ is interpreted as "agent i views agent j as an agent with high credibility about φ". If we suppose that each agent's perspective on credibility is correct and that these viewpoints are common knowledge among agents, then we can consider those notions concerning highness, neutralness, and lowness as an order relation. In other words, the axioms which we are going to describe are intended to describe the situation where credibility is directly linked to the true observable expertise of the agents. This expertise moreover should be common knowledge among agents. For this scenario, we have the following axioms for our credibility operators:

Axioms:

Irreflexivity $\neg HL_{i,i}\varphi$.
Asymmetry $HL_{i,j}\varphi \rightarrow \neg HL_{j,i}\varphi$.
Transitivity $HL_{i,j}\varphi \wedge HL_{j,k}\varphi \rightarrow HL_{i,k}\varphi$.
Reflexivity $NL_{i,i}\varphi$.
Symmetry $NL_{i,j}\varphi \rightarrow NL_{j,i}\varphi$.

Transitivity $NL_{i,j}\varphi \wedge NL_{j,k}\varphi \rightarrow NL_{i,k}\varphi$.

Definition $LL_{i,j}\varphi \stackrel{\text{def}}{\Longleftrightarrow} HL_{j,i}\varphi$.

A little reflection shows that the above axioms can easily be violated in situations where agents are misinformed about each other's expertise, or where there exists disagreements among the agents.

Next we make some analysis on possible configurations among the relied-on relations. In order to obtain an intuitive understanding on this approach, we will use the notion of *role* to refer to different types of agents characterized in terms of being relied on in different configurations; this relates to role theory in social science [Biddle 1979, Jackson 1972].

For a single agent, there exist the following different roles which the agent can perform (for the formula φ):

Isolated-A The agent relies on nobody including himself and is relied on by nobody: $(\neg\exists j)(D_{i,j}\varphi \vee D_{j,i}\varphi)$
Isolated-B The agent relies on only himself, while nobody relies on him: $D_{i,i}\varphi \wedge (\neg\exists j \neq i)(D_{j,i}\varphi \vee D_{i,j}\varphi)$
Learner The agent only relies on someone else: $(\exists j \neq i)D_{i,j}\varphi \wedge \neg D_{i,i}\varphi$.
Expert The agent relies on both himself and someone else: $D_{i,i}\varphi \wedge (\exists j \neq i)D_{i,j}\varphi$
Authority The agent is relied on both by himself and by someone else, but he relies on nobody else: $D_{i,i}\varphi \wedge (\exists j \neq i)D_{j,i}\varphi \wedge \neg(\exists k \neq i)D_{i,k}\varphi$.

Diffident agent The agent relies on nobody including himself, but is relied on by someone else: $(\exists j \neq i)D_{j,i}\varphi \wedge \neg(\exists k)D_{i,k}\varphi$

Among the above roles, Role Isolated-A and Role Isolated-B are isolated ones where agents rely on nobody else; in the studies of belief dependence these are trivial roles. The remaining roles are the fundamental roles which are worthy to be named and investigated in depth. For these roles we introduce the notations given below:

$Learner_i\varphi \stackrel{\text{def}}{\Longleftrightarrow} \neg D_{i,i}\varphi \wedge (\exists j \neq i)D_{i,j}\varphi$.
$Expert_i\varphi \stackrel{\text{def}}{\Longleftrightarrow} D_{i,i}\varphi \wedge (\exists j \neq i)D_{i,j}\varphi$.
$Authority_i\varphi \stackrel{\text{def}}{\Longleftrightarrow} D_{i,i}\varphi \wedge (\exists j \neq i)D_{j,i}\varphi \wedge \neg(\exists k \neq i)D_{i,k}\varphi$.
$Diffident-agent_i\varphi \stackrel{\text{def}}{\Longleftrightarrow} (\exists j \neq i)D_{j,i}\varphi \wedge \neg(\exists k)D_{i,k}\varphi$.

5.5.3 Roles and Credibilities

Using the fundamental roles described above, we now propose a strategy, called the *confidence priority strategy*, yielding an intuitive mechanism for choosing between the three credibility operators. This strategy enforces that the more confidently a formula φ is believed by an informer, the more credible the belief is. We believe that the confidence priority strategy is reasonable and acceptable for cooperative multiple agent environments. If an agent firmly believes a fact φ, to the extent that others rely on him then this indicates that the agent must have strong evidence or a convincing justification for his belief. The other agents therefore should view the agent's belief as a belief with higher credibility.

According to the confidence priority strategy, the fundamental roles can be arranged intuitively in increasing order as diffident-agent, learner, expert, and authority. Note that in our formalism the notion of credibility is a relative one. For instances, an agents who is a learner will view agents that are expert or authority as agents with high credibility about φ. Whereas neutral credibility will be granted to peers, i.e. in the case where both agents are learners, experts, or authorities. The relative credibility relations can be formalized as follows:

Def-High-credibility $HL_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} (Diffident-agent_i\varphi \wedge (Learner_j\varphi \vee Expert_j\varphi \vee Authority_j\varphi)) \vee (Learner_i\varphi \wedge (Expert_j\varphi \vee Authority_j\varphi)) \vee (Expert_i\varphi \wedge Authority_j\varphi)$.

Def-Neutral-credibility $NL_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} (Diffident-agent_i\varphi \wedge Diffident-agent_j\varphi) \vee (Learner_i\varphi \wedge Learner_j\varphi) \vee (Expert_i \wedge Expert_j) \vee (Authority_i \wedge Authority_j\varphi)$.

Def-Low-credibility $LL_{j,i}\varphi \overset{\text{def}}{\Longleftrightarrow} HL_{i,j}\varphi$.

If we combine our proposals in this section and the previous one we obtain a computational strategy for the standard case of the belief conflict under consideration, i.e. the situation where $\varphi \in L^{-}_{i,j}(K) \wedge \neg\varphi \in L_i(K)$. For the three meaningful revision operators we will use the notations **Positive-revision**, **Persistence**, and **Negative-contraction** to denoted the corresponding processes.

If we suppose that axiom (Ldf) $L_{i,j}\varphi \overset{\text{def}}{\Longleftrightarrow} D_{i,j}\varphi \wedge L_j\varphi$ and the neutral axiom (D$\neg$) $D_{i,j}\varphi \leftrightarrow D_{i,j}\neg\varphi$ hold, we have, for the case of normal belief conflict that $\varphi \in L^{-}_{i,j}(K) \wedge \neg\varphi \in L_i(K)$ implies $D_{i,j}\varphi \wedge L_j\varphi \wedge L_i\neg\varphi$ Therefore, we also call the case that $D_{i,j}\varphi \wedge L_j\varphi \wedge L_i\neg\varphi$ a normal belief conflict.

For this case of normal belief conflict, we have:

$D_{i,j}\varphi \wedge L_j\varphi \wedge L_i\neg\varphi \Rightarrow D_{i,j}\varphi \Rightarrow (D_{i,j}\varphi \vee D_{i,i}\varphi) \wedge (D_{i,j}\varphi \vee \neg D_{i,i}\varphi) \Rightarrow Expert_i\varphi \vee Learner_i\varphi$.

In the case that agent i is an expert about φ, and if $\neg D_{j,j}\varphi$ holds, it follows that agent j is a learner or a diffident-agent, and we have $LL_{i,j}\varphi$. Therefore, the outcome will be persistence, denoted by $\sim$ for short. On the other hand, if $D_{j,j}\varphi$ holds, and if there exists an agent k such that $D_{j,k}\varphi$, the agents j and i are both experts and therefore peers and the outcome will be negative-contraction. Otherwise the process

would be positive-revision because agent j is an authority about φ.

For the other case that agent i is a learner about φ, we investigate once more whether $D_{j,j}\varphi$ holds. If this is the case that means that agent j at least is an expert and we have $HL_{i,j}\varphi$ and the process will be positive-revision. On the other hand, if $\neg D_{j,j}\varphi \wedge (\exists k \neq j)D_{j,k}\varphi$ holds, then agent j is a learner, and the outcome will be negative-contraction. In the remaining case the outcome is persistence.

Therefore, the strategy can be expressed as follows:

$D_{i,j}\varphi \wedge L_j\varphi \wedge L_i\neg\varphi \wedge D_{i,i}\varphi \wedge \neg D_{j,j}\varphi \Rightarrow$**Persistence**.

$D_{i,j}\varphi \wedge L_j\varphi \wedge L_i\neg\varphi \wedge D_{i,i}\varphi \wedge D_{j,j}\varphi \wedge (\exists k \neq j)D_{j,k}\varphi \Rightarrow$**Negative-contraction**.

$D_{i,j}\varphi \wedge L_j\varphi \wedge L_i\neg\varphi \wedge D_{j,j}\varphi \wedge \neg(\exists j \neq k)D_{j,k}\varphi \Rightarrow$**Positive-revision**.

$D_{i,j}\varphi \wedge L_j\varphi \wedge L_i\neg\varphi \wedge \neg D_{i,i}\varphi \wedge D_{j,j}\varphi \Rightarrow$**Positive-revision**.

$D_{i,j}\varphi \wedge L_j\varphi \wedge L_i\neg\varphi \wedge \neg D_{i,i}\varphi \wedge D_{j,j}\varphi \wedge (\exists k \neq j)D_{j,k}\varphi \Rightarrow$**Negative-contraction**.

$D_{i,j}\varphi \wedge L_j\varphi \wedge L_i\neg\varphi \wedge \neg D_{i,i}\varphi \wedge \neg D_{j,j}\varphi \wedge \neg(\exists k \neq j)D_{j,k}\varphi \Rightarrow$**Persistence**.

Moreover, the strategy can be expressed as a decision tree as it is shown in the figure 5.1.

5.6 Conclusions

Based on the framework of logic for belief dependence, we have proposed a rational balance strategy between belief revision and belief persistence in a multiple agent environment. The confidence priority strategy provides an intuitive, plausible and flexible approach to formalize the relationship between the fundamental roles and credibility operators.

On the other hand, we are convinced that there exist many other strategies which are also plausible and acceptable for multiple agent environments; such alternatives could be based on different analysis and perspectives on the rely-on relations in the models. To capture the other plausible strategies is an interesting topic for further research.

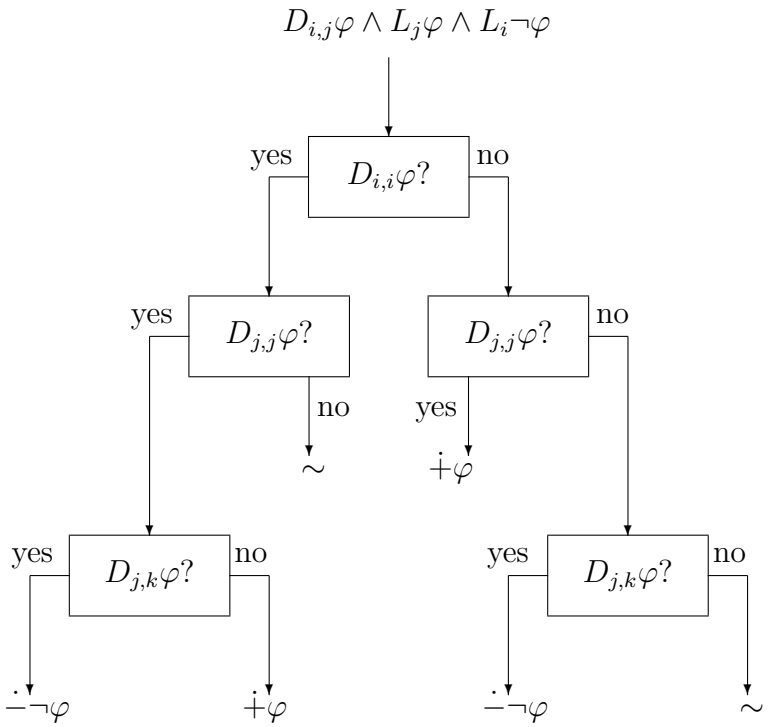


Figure 5.1: Decision Tree

Chapter 6

Information Acquisition from a Multi-agent Environment

6.1 Schoenmakers Problem

The construction of models for multi-agent epistemic systems has become one of the most interesting and popular topics in artificial intelligence and in the theory of knowledge based expert systems. Information systems in the real world are loaded by combining information from many (possibly unrelated) sources. As is generally known merging information may produce inconsistent knowledge bases. However, an even more subtle risk was indicated by W. J. Schoenmakers[Schoenmakers 1986] when he published his *Judge puzzle*. This puzzle describes the situation where an agent, called the *judge*, by combining information from two other agents, called the *witnesses*, consistently obtains a conclusion which contradicts the belief sets of both witnesses:

> *Once upon a time a wise but strictly formal judge questioned two witnesses. They spoke to her on separate occasions. Witness* $w1$ *honestly stated his conviction that proposition* p *was true. Witness* $w2$ *honestly stated that he believed that the implication* $p \rightarrow q$ *was true. Nothing else was said or heard. The judge, not noticing any inconsistency accepted both statements and concluded that* q *had to be true. However, when the two witnesses heard about her conclusion they were shocked because they both were convinced that* q *was false. But they were too late to prevent the verdict to be executed...*

As pointed out by Schoenmakers, in the above story nobody can be blamed for this situation to arise. The witnesses, even though formally required to tell everything they know, are not responsible since neither of them was asked about q and hardly could know at the time of interrogation that the truth of q was at stake. The judge on the other hand had no reason to even consider the possibility that her argument was unsound, since there is not the slightest trace of contradiction

in the testimony. She might have asked on, and confronted the witnesses with her conclusion that q was true. For the judge this would have been possible, but, as Schoenmakers indicates, this possibility is lost in the case of a knowledge base being loaded with information from independent sources, since by the time proposition q turns out to be relevant the two informants no longer are accessible. And therefore Schoenmakers concludes:

> *Intelligent database systems may behave perfectly in splendid isolation, operating on one world without inconsistencies, but even when they are consistent they may produce unacceptable results when operating on the information that is accessible in a community of such systems. Their results will be acceptable, most of the time, but nobody knows when.*

Consequently it becomes relevant to look for a characterization of situations where combining information from multi sources is *safe*, which informally means that no conclusion drawn from the combined information is disbelieved by all informants. At the same time our combination operator should support at least the derivation of one proposition not already supported by one of the contributing agents; otherwise the problem of obtaining the right information reduces to the identification of the right source.

However, having formalized this problem, we encounter a triviality theorem expressing that a combining operation satisfying the above form of *absolute safety* doesn't exist. Consequently, a more refined approach is required which takes into account both the information contributed by the agents and their complete belief sets. In this context the notions of *safety* and *strong safety* are defined, and some characterizations are obtained. It follows that dangerous situations only arise when every agent disagrees with some other agent about some of the propositions which are actually communicated.

These results once more indicate that in a multi agent environment one should maintain a strict distinction between information accepted on behalf of an other agent, and information which is incorporated in your own belief set. The resulting process of accepting information followed by incorporating it, as is argued before, is one of the main motivations for the introduction of the logic of belief dependence.

The triviality theorem shows that there is no simple solution for the problem. The characterization of the less restrictive safety notions shows that danger is caused by disagreement between agents and disagreement between agents is a fact of life we can't get around. The case for a two stage process for belief incorporation has been argued in the previous chapters; it is also supported by psychological research. However, when generalizing the safety notions to the case of our epistemic logic of belief dependence, the characterizations for the propositional case extend, and so do their negative consequences. Therefore, the best we can hope for is a specific belief incorporation strategy for the judge which is approximatively safe.

The proposed notion of *restricted almost safety* characterizes the situation where the conclusion of the judge will not be contradicted by all witnesses, provided they will eventually have access to each other's information. This hypothetical situation can be expressed in terms of sub-beliefs in our logic of belief dependence, leading

to an effectively testable condition for deciding whether a specific belief revision operator for the judge is almost safe or not.

6.2 Combining Information from Multiple Agents; the triviality result

In the sequel I denotes a finite and non-empty set of the agents called *informants* and a the *receiver*, an agent who receives and combines information from the informants I. In this section, we study the case of propositional logic $\mathbf{L_P}$, where information communicated between agents consists of pure propositional formulas without modal operators. The language $\mathbf{L_P}$ is recursively constructed from a primitive proposition set Φ_0 and the Boolean connectives as usual. Moreover, the logical notions of a semantic model, the satisfiability relations $\models$, and the consequence operation Cn, are defined as usual.

The receiver's obtained information, is a mapping ψ from the informants I into the formula set $\mathbf{L_P}$. We use the notation $\{\psi_i\}_{i\in I}$ to denote the set $\{\psi(i) \in \mathbf{L_P} : i \in I\}$. The set $\{\psi_i\}_{i\in I}$ is called the *obtained information set*. Each informant may contribute a finite set of formulas which expresses his share in the information exchange; this finite set clearly can be reduced to a single formula by taking the corresponding conjunction formula.[1] Furthermore, the informants' original belief sets are represented by a mapping Ψ from the informant set I into the powerset of the formula set. We also use the notation $\{\Psi_i\}_{i\in I}$ to denote the set $\{\Psi(i) \in \mathcal{P}(\mathbf{L_P}) : i \in I\}$, which is called an *original information set*. These sets $\{\Psi_i\}_{i\in I}$ are not required to be finite. In this thesis, we only consider the case where all informants honestly offer information they actually support. This leads to the following definition:

6.2.1. Definition. (Potential information set) *An original information set $\{\Psi_i\}_{i\in I}$ is said to be a potential information set of an obtained information set $\{\psi_i\}_{i\in I}$ iff it satisfies the following conditions:*
(i) (Honesty Condition) $\Psi(i) \models \psi(i)$, for all $i \in I$, and
(ii) (Consistency Condition) $\Psi(i)$ is consistent, for all $i \in I$.

In the sequel we shall use the word set for information set when no confusion can arise.

6.2.2. Definition. (Danger) *Suppose that some original set $\{\Psi_i\}_{i\in I}$ is a potential set of an obtained set $\{\psi_i\}_{i\in I}$. Then the set $\{\psi_i\}_{i\in I}$ is said to be* dangerous *with respect to the set $\{\Psi_i\}_{i\in I}$ iff there exists a $\varphi \in \mathbf{L_P}$ such that*
(i) $\{\psi_i\}_{i\in I} \models \varphi$,
(ii) $\Psi(i) \models \neg\varphi$ for all $i \in I$.

Remarks: Condition (i) means that the receiver's obtained information implies some fact φ for which according to condition (ii) all informants originally believe its

[1] Here we use the fact that the languages considered in this paper are closed under conjunction; the case where we don't assume this closure property is a subject for further research.

negation. The more general notion where some derivable fact φ is disbelieved by some but not necessarily all informants is not interesting for our purposes; a contributed set will be "dangerous" in this more general sense with respect to an original set, unless it represents a proposition which is already compatible with the original belief set of all informants. The latter situation is frequently considered in artificial intelligence, where collected information always represent a partial description of the true world. In our approach we don't require such a true world in the background; we just want to ensure that derivable information is at least compatible with the beliefs of some agent.
In the following, $\{\psi_i\}_{i\in I}$ and $\{\Psi_i\}_{i\in I}$ denote an obtained set and an original set respectively if this does not cause ambiguities.

6.2.3. Definition. (Absolute safety) *A consistent set $\{\psi_i\}_{i\in I}$ is said to be* absolutely safe *iff it is not the case that $\{\psi_i\}_{i\in I}$ is dangerous with respect to any of its potential sets $\{\Psi_i\}_{i\in I}$.*

6.2.4. Definition. (Triviality) *A set $\{\psi_i\}_{i\in I}$ is trivial iff for any formula φ, such that $\{\psi_i\}_{i\in I} \models \varphi$, there exists an $i \in I$ such that $\psi(i) \models \varphi$.*

Clearly a set $\{\psi_i\}_{i\in I}$ is trivial iff some formula $\psi(i)$ is logically equivalent to $\bigwedge\{\psi_i\}_{i\in I}$, which means that in fact one informant has already contributed all available information by himself. This observation easily follows by taking $\varphi = \bigwedge\{\psi_i\}_{i\in I}$. It turns out that absolute safety is a condition which is so strong that it supports only trivial situations:

6.2.5. Theorem. (Triviality theorem) *A consistent set $\{\psi_i\}_{i\in I}$ is absolutely safe iff it is trivial.*

PROOF. The proof for this result is easy. Assuming non-triviality there exists a proposition $\phi(i)$ such that for no i one has $\psi \models \varphi$; consequently the potential set $\Psi(i) = \{\psi(i), \neg\varphi\}$, for all $i \in I$ is dangerous with respect to $\{\psi_i\}_{i\in I}$. The converse implication is a direct consequence of the triviality condition. □

The triviality theorem tells us that one cannot have his or her obtained multi resource information being both absolutely safe and non-trivial. The theorem implies that seeking absolutely safe information is not practical. More reasonably, we should say the information is safe with respect to a specified original set rather than with respect to all potentially original sets. However, the notion of safety with respect to a specified original set requires the complete information about the informants' original belief sets, which seems to be not realistic, because, in the applications, a rational agent normally has no such complete knowledge or belief. However, by the new notion of safety, as shown below, we can find some interesting and useful properties, which offers a better understanding of the Schoenmakers problem.

6.2.6. Definition. (Safety) *If an obtained $\{\psi_i\}_{i\in I}$ is consistent, and an original set $\{\Psi_i\}_{i\in I}$ is a potential set of $\{\psi_i\}_{i\in I}$, then the set $\{\psi_i\}_{i\in I}$ is said to be* safe *with respect to the set $\{\Psi_i\}_{i\in I}$ iff the set $\{\psi_i\}_{i\in I}$ is not dangerous with respect to the set $\{\Psi_i\}_{i\in I}$.*

It is easy to see that we have the following proposition about the safety, which is straightforward from the definition.

6.2.7. PROPOSITION. *If $\{\psi_i\}_{i\in I}$ is a consistent set, and $\{\Psi_i\}_{i\in I}$ is a potential set of $\{\psi_i\}_{i\in I}$, then the set $\{\psi_i\}_{i\in I}$ is safe with respect to its potential set $\{\Psi_i\}_{i\in I}$ iff*
for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then it is not the case that for all $i \in I$, $\Psi(i) \models \neg\varphi$.

The above proposition about the safety suggests an alternative to define a stronger notion about the safety.

6.2.8. DEFINITION. (Strong safety) *If a set $\{\psi_i\}_{i\in I}$ is consistent, and $\{\Psi_i\}_{i\in I}$ is a potential set of $\{\psi_i\}_{i\in I}$, then the set $\{\psi_i\}_{i\in I}$ is said to be* strongly safe *with respect to the set $\{\Psi_i\}_{i\in I}$ iff*
for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then there exists an $i \in I$ such that $\Psi(i) \models \varphi$.

So where the safety condition requires that every derivable formula φ is not disbelieved by all informants, the condition of strong safety requires that at least one of the informants positively supports φ. It follows that strong safety is a stronger notion than safety. Suppose that we have a linear ordering $\preceq$ on the informant set $I = \{i_1, i_2, ..., i_m\}$, say, $i_1 \preceq i_2 \preceq i_3 \preceq ... \preceq i_m$. We can use a tuple $\langle\psi(i_1), \psi(i_2), ..., \psi(i_m)\rangle$ to denote the set $\{\psi_i\}_{i\in I}$. Also, we can use a tuple $\langle\Psi(i_1), \Psi(i_2), ..., \Psi(i_m)\rangle$ to denote the set $\{\Psi_i\}_{i\in I}$. Here are some examples about the safety and the strong safety.

6.2.9. EXAMPLE.
- *$\langle p, p \to q\rangle$ is neither strongly safe nor safe with respect to the potential set*
  $\langle\{p, \neg q\}, \{\neg p, \neg q\}\rangle$.(**Judge puzzle**)
- *$\langle p \to q, q \to p\rangle$ is strongly safe and safe with respect to $\langle\{\neg p, q\}, \{\neg p, \neg q\}\rangle$.*
- *$\langle p, p \to q\rangle$ is safe with respect to $\langle\{p, p \vee q\}, \{p \to q, q \to p\}\rangle$, but not strongly safe with respect to $\langle\{p, p \vee q\}, \{p \to q, q \to p\}\rangle$.* (**Distinction between safety and strong safety**)

6.2.10. PROPOSITION. *If a consistent set $\{\psi_i\}_{i\in I}$ is trivial, then the set $\{\psi_i\}_{i\in I}$ is strongly safe with respect to any of its potential set $\{\Psi_i\}_{i\in I}$.*

PROOF. Suppose that $\{\psi_i\}_{i\in I}$ is trivial. Thus, for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then there exists an $i \in I$ such that $\psi(i) \models \varphi$, by the definition of the triviality. Therefore, for any φ, if $\psi(i) \models \varphi$, then for any $\Psi(i)$ such that $\Psi(i) \models \psi(i)$, we have that $\Psi(i) \models \varphi$. So, by the definition of potential set, for any potential set $\{\Psi_i\}_{i\in I}$ of $\{\psi_i\}_{i\in I}$, we have $\Psi(i) \models \varphi$. So $\{\psi_i\}_{i\in I}$ is strongly safe with respect to any of its potential set $\{\Psi_i\}_{i\in I}$. □

Because the informant set I is finite, the set $\{\psi_i\}_{i\in I}$ also is a finite formula set. We use the notation $\bigwedge\{\psi_i\}_{i\in I}$ to denote the conjunction formula of the set $\{\psi_i\}_{i\in I}$.

6.2.11. THEOREM. (Safety theorem) *If an obtained set $\{\psi_i\}_{i\in I}$ is consistent, and $\{\Psi_i\}_{i\in I}$ is a potential set of $\{\psi_i\}_{i\in I}$, then the set $\{\psi_i\}_{i\in I}$ is safe with respect to the original set $\{\Psi_i\}_{i\in I}$ iff there exists an $i \in I$ such that $\Psi(i) \cup \{\psi_i\}_{i\in I}$ is consistent.*

PROOF. ($\Rightarrow$) Suppose that $\{\psi_i\}_{i\in I}$ is safe with respect to $\{\Psi_i\}_{i\in I}$. Thus, by the definition, for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then there exists an $i \in I$ such that $\Psi(i) \not\models \neg\varphi$. Specially, we have $\{\psi_i\}_{i\in I} \models \bigwedge\{\psi_i\}_{i\in I}$. Therefore, there exists an $i \in I$ such that $\Psi(i) \not\models \neg(\wedge\{\psi_i\}_{i\in I})$. So $\Psi(i) \cup \{\bigwedge\{\psi_i\}_{i\in I}\}$ is consistent, and consequently $\Psi(i) \cup \{\psi_i\}_{i\in I}$ is consistent.

($\Leftarrow$) Suppose that there exists an $i \in I$ such that $\Psi(i) \cup \{\psi_i\}_{i\in I}$ is consistent. Thus, for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then $\Psi(i) \cup \{\varphi\}$ is consistent. So $\Psi(i) \not\models \neg\varphi$. This proves that $\{\psi_i\}_{i\in I}$ is safe with respect to $\{\Psi_i\}_{i\in I}$. □

6.2.12. Lemma. (Multi-informants lemma) *If the informant set is a singleton, i.e., $|I| = 1$, then any consistent set $\{\psi_i\}_{i\in I}$ is absolutely safe. Therefore, $\{\psi_i\}_{i\in I}$ is safe with respect to any of its potential set $\{\Psi_i\}_{i\in I}$.*

PROOF. Straightforward from the definition of safety. □

6.2.13. Lemma. *If a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$, then there exists a formula φ such that $\{\psi_i\}_{i\in I} \models \varphi$ and for all $i \in I$, $\psi(i) \not\models \varphi$ and $\Psi(i) \models \neg\varphi$.*

PROOF. Suppose that a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$. Then by the definition, there exists a φ such that $\{\psi_i\}_{i\in I} \models \varphi$ and $\Psi(i) \models \neg\varphi$ for all $i \in I$. It is easy to see that for any $i \in I$, we have $\psi(i) \not\models \varphi$, because, if $\psi(i) \models \varphi$, then $\Psi(i) \models \varphi$, by the definition of the potential set. However, any $\Psi(i)$ in the potential set is consistent, this contradicts $\Psi(i) \models \neg\varphi$. □

6.2.14. Lemma. *If a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$, then for all $i \in I$, $\psi(i) \not\models \bigwedge\{\psi_i\}_{i\in I}$ and $\Psi(i) \models \neg\bigwedge\{\psi_i\}_{i\in I}$.*

PROOF. Suppose that a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$. Then, by definition, there exists a φ such that $\{\psi_i\}_{i\in I} \models \varphi$ and $\Psi(i) \models \neg\varphi$ for all $i \in I$. Therefore, $\bigwedge\{\psi_i\}_{i\in I} \models \varphi$, and consequently $\models \bigwedge\{\psi_i\}_{i\in I} \rightarrow \varphi$ and by contraposition $\models \neg\varphi \rightarrow \neg\bigwedge\{\psi_i\}_{i\in I}$.
However, since $\{\psi_i\}_{i\in I}$ is dangerous $\Psi(i) \models \neg\varphi$ for any $i \in I$ and therefore, $\Psi(i) \models \neg\bigwedge\{\psi_i\}_{i\in I}$ for any $i \in I$.
Finally, it is easy to see that for any $i \in I$, $\psi(i) \not\models \bigwedge\{\psi_i\}_{i\in I}$, because, if $\psi(i) \models \bigwedge\{\psi_i\}_{i\in I}$ for any $i \in I$, then $\{\psi_i\}_{i\in I}$ is trivial, and then, by the triviality theorem, $\{\psi_i\}_{i\in I}$ is absolutely safe, whence $\{\psi_i\}_{i\in I}$ cannot be dangerous with respect to any potential set, and a contradiction follows. □

6.2.15. THEOREM. (Disagreement theorem) *If a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$, then there exists for every $j \in I$ some formula φ and an $i \in I$ such that $\psi(i) \models \varphi$ and $\Psi(j) \not\models \varphi$.*

PROOF. Suppose that a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$. Then by the above lemma, we have,

(A) $\Psi(j) \models \neg \bigwedge\{\psi_i\}_{i\in I}$ for all $j \in I$.

Now, suppose that the conclusion (B) of the disagreement theorem is false, then we have (C).

(B) $(\forall j \in I)(\exists\varphi)(\exists i \in I)(\psi(i) \models \varphi$ and $\Psi(j) \not\models \varphi)$.
(C) $(\exists j \in I)(\forall\varphi)(\forall i \in I)(\psi(i) \models \varphi \Rightarrow \Psi(j) \models \varphi)$.

However, we know that $\psi(i) \models \psi(i)$ for any $i \in I$. Therefore, by (C), we have,
$(\exists j \in I)(\forall i \in I)(\Psi(j) \models \psi(i))$.
So, we have,

(D) $(\exists j \in I)(\Psi(j) \models \bigwedge\{\psi_i\}_{i\in I})$.

Consequently, from (D) and (A), we conclude that this particular $\Psi(j)$ is inconsistent, contradicting our assumption that $\{\Psi_i\}_{i\in I}$ is a potential set. □

Observe that the conclusion of the disagreement theorem can't be strengthened to a form which expresses definite disagreement: *there exists for every $j \in I$ some formula φ and an $i \in I$ such that $\psi(i) \models \varphi$ and $\Psi(j) \models \neg\varphi$.* This conclusion only can be proven if we assume that the sets $\{\Psi_i\}_{i\in I}$ satisfy the strong condition $\Psi(j) \models (\phi \vee \psi)$ iff $\Psi(j) \models \phi$ or $\Psi(j) \models \psi$, i.e., if we assume that our agents use an intuitionistic interpretation of disjunction.

6.2.16. COROLLARY. *If a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$, then there exist for every $j \in I$ an $i \in I$ and a formula φ in the consequence set $Cn(\{\psi(i)\})$ such that $\Psi(i) \models \varphi$ and $\Psi(j) \not\models \varphi$; such a formula φ is called a* disagreement formula *for j.*

In the judge puzzle story, the formula $p \to q$ is a disagreement formula for w_1, since $\Psi(w_1) = \{p, \neg q\} \not\models p \to q$ and $\Psi(w_2) = \{\neg p, \neg q\} \models p \to q$.

The implication of the disagreement theorem is that in a multi-agent information system in order to guarantee safety, agents must be prohibited to talk about something if they disagree with someone else about it! Therefore each agent will need full information about the others' propositional attitudes, and this clearly represents an unrealistic assumption. Nonetheless the result implies that we should focus on the

cases where disagreement may arise, and look for mechanisms for coping with it. As we indicate below, the logic of belief dependence turns out to be a useful tool in this direction.

6.3 Information Acquisition in a Belief Dependence Framework

In this section, we consider the information acquisition problem in our framework of belief dependence logic. We use the logic of belief dependence $\mathbf{LD}^+$ to study the problem. The extension to the definitions which have appeared in section 2 for the case of belief dependence is easy: one simply replaces the propositional language $\mathbf{L_P}$ by the language $\mathbf{L_D}$, propositional models by D-models, and the relation $\models$ for propositional logic by its counterpart for belief dependence logic. Consequently, whenever we say a formula set K is consistent, we mean that K is consistent with respect to the $\mathbf{LD}^+$ system unless stated otherwise.

In the resulting theory the (negative) results from section 2 remain valid, indicating that for a solution of problems like the Judge puzzle the formalization of the relevant information into the language of belief dependence logic by itself will be insufficient in order to remove the observed anomaly.

The translation between the propositional formulation of our problem and its formalization in terms of belief states K in the logic of belief dependence invokes a few auxiliary notations defined below:

Recall that $L^-_{i,j}(K) \stackrel{\text{def}}{\Longleftrightarrow} \{\varphi \in \mathbf{L_D} : K \models L_{i,j}\varphi\}$, denotes agent i's compartmentalized belief set indexed j, and $L^-_i(K) \stackrel{\text{def}}{\Longleftrightarrow} \{\varphi \in \mathbf{L_D} : K \models L_i\varphi\}$, denotes agent i's (incorporated) belief set.

Moreover, we define that $L^+_{a,I}(\{\psi_i\}_{i\in I}) \stackrel{\text{def}}{\Longleftrightarrow} \{L_{a,i}\psi(i) \in \mathbf{L_D} : i \in I\}$, is a formula set expressing the information obtained by the receiver from the informants before the receiver has incorporated (part of) this information.

The notion of a *configuration* represents a generalization of a potential set from section 2:

6.3.1. Definition. (Configuration) *A configuration C is a tuple $\langle a, I, \psi, \Psi\rangle$, where $a \in A_n$ denotes an agent, called receiver, $I \subseteq A_n$ is a finite and non-empty set of informants, $\psi : I \to \mathbf{L_D}$ is a mapping from I into $\mathbf{L_D}$, called the obtained information, and $\Psi : I \to \mathcal{P}(\mathbf{L_D})$ is a mapping from I into the powerset of $\mathbf{L_D}$, called the original information.*

Since the required proofs are insensitive to the precise logical language being used it will not surprising that the main results of section 2 remain valid for the logic of belief dependence:

6.3.2. Theorem. (Triviality theorem (restated)) *A consistent obtained set $\{\psi_i\}_{i\in I}$ is absolutely safe iff it is trivial.*

6.3.3. Theorem. (Disagreement theorem (restated)) *Let $C = \langle a, I, \psi, \Psi\rangle$ be a configuration. Suppose that $\{\psi_i\}_{i\in I}$ is consistent, and $\{\Psi_i\}_{i\in I}$ is a potential set of*

$\{\psi_i\}_{i\in I}$. If $\{\psi_i\}_{i\in I}$ is dangerous with respect to the set $\{\Psi_i\}_{i\in I}$, then there exists for every agent $j \in I$ a formula φ and agent $i \in I$ such that $\psi(i) \models \varphi$ and $\Psi(j) \not\models \varphi$.

For a belief state K in belief dependence logic and an agent a, we want to induce a configuration for a from K. In this induced configuration agent a becomes the receiver and the remaining agents become the informants. Both the contributed information and the original information is obtained from the belief set K as indicated below.

6.3.4. Definition. (Induced configuration) *Suppose that K be a belief state, and a be an agent $\in A_n$. A configuration $C = \langle a, I, \psi, \Psi\rangle$, called the induced configuration for a from K, is constructed as follows:*
(1) I is the set $\{i \in A_n : \exists\varphi(L_{a,i}\varphi \in K)\}$.
(2) If I is not empty, then for all $i \in I, \Psi(i) = L_i^-(K)$, otherwise the induced configuration does not exist.
(3) For all $i \in I$, if $L_{a,i}^-(K)$ is finite, then let $\psi(i)$ be $\wedge L_{a,i}^-(K)$, otherwise the induced configuration does not exist.

For an agent $a \in A_n$ a belief state K is said to be a *DB set for a* iff the induced configuration for a from K exists. Evidently the induced configuration $\langle a, I, \psi, \Psi\rangle$ for a from K is unique whenever it exists. We introduce the notation $C(a, K)$ for the induced configuration for a from K. Moreover, due to the honesty condition contained in definition (Lijdf) the original information set $\Psi(i)$ is a potential set for $\psi(i)$ for each $i \in I$. The concept of the induced configuration makes it possible to translate the safety definitions from section 2 to belief states in belief dependence logic:

6.3.5. Definition. (Safety for a in K) *For an agent $a \in A_n$ and a DB set K for a, let $C(a, K) = \langle a, I, \psi, \Psi\rangle$ be the induced configuration for a from K, then $\{\psi_i\}_{i\in I}$ is said to be safe for a in K iff $\{\psi_i\}_{i\in I}$ is safe with respect to $\{\Psi_i\}_{i\in I}$.*

6.3.6. Lemma. (Safety lemma) *For an agent $a \in A_n$ and a DB set K for a, if $C(a, K) = \langle a, I, \psi, \Psi\rangle$ is the induced configuration for a from K.*
$\{\psi_i\}_{i\in I}$ is safe for a in K iff
for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then there exist $i \in I$ such that $L_i^-(K) \not\models \neg\varphi$

6.3.7. Theorem. (Safety theorem (restated)) *Let a and K be an agent and a DB set respectively. Suppose that the induced configuration for a from K, $C(a, K) = \langle a, I, \psi, \Psi\rangle$.*

$\{\psi_i\}_{i\in I}$ is safe for a in K iff there exists an $i \in I$ such that $L_i^-(K) \cup \{\psi_i\}_{i\in I}$ is consistent.

PROOF. The theorem is a natural generalization from the safety theorem concerning propositional logic.

($\Rightarrow$) Suppose that $\{\psi_i\}_{i\in I}$ is safe for a in K. Thus, by the safety lemma, for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then there exists a i in I such that $L_i^-(K) \not\models \neg\varphi$. Specially, we have

$\{\psi_i\}_{i\in I} \models \wedge\{\psi_i\}$. Therefore, there exist $i \in I$ such that $L_i^-(K) \not\models \neg(\wedge\{\psi_i\}_{i\in I})$. So $L_i^-(K) \cup (\wedge\{\psi_i\}_{i\in I})$ is consistent.
($\Leftarrow$) Suppose that there exist $i \in I$ such that $L_i^-(K) \not\models \neg(\wedge\{\psi_i\}_{i\in I})$. Thus, for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then $L_i^-(K) \cup \{\varphi\}$ is consistent and $L_i^-(K) \not\models \neg\varphi$. Therefore, $\{\psi_i\}_{i\in I}$ is almost safe for a on K. □

6.4 Almost Safety

In order to evaluate whether obtained information is safe the receiver a still needs information on the true belief states of his informants; the translation into the belief dependence logic and the introduction of configurations has not changed this necessity. However, if we take into consideration which mechanisms might have produced the sub-beliefs in a multi-agent environment, it turns out that these mechanisms themselves may provide us with additional structure supporting the introduction of alternative and weaker safety notions.

The notion of *almost safety* defined in this section is based on one possible hypothesis concerning the creation of sub-beliefs: the so-called *strong initial role-information assumption*. This hypothesis states that within a multi-agent environment the dependency relations are common knowledge: it is not known who knows what or who believes what, but for each proposition it is known how the agents depend on each other concerning this proposition.

That this information is relevant is shown by the example below. Assume that some agent i believes ϕ and says so to the receiver. Suppose moreover that the receiver has learned previously that agent j believes $\neg\phi$. Finally agent i depends on agent j concerning ϕ. According to the strong initial role-information assumption it is common knowledge that $D_{i,j}\phi$, so the receiver knows that as well. In this situation the receiver can conclude that something strange is going on: would the two agents i and j have been given the possibility to exchange their information, agent i would have been convinced by j that his belief concerning ϕ was wrong. Moreover, this prediction can be made by the receiver without any further interaction with the informants! It is based on this information that the receiver can disregard the information provided by i substituting it by the opposite information provided by agent j.

The notion of almost safety formalizes safety with respect to the hypothetical scenario which will arise when all informants exchange their information before sharing their knowledge with the receiver. In order to be able to reason about these hypothetical belief states we need one further notion:

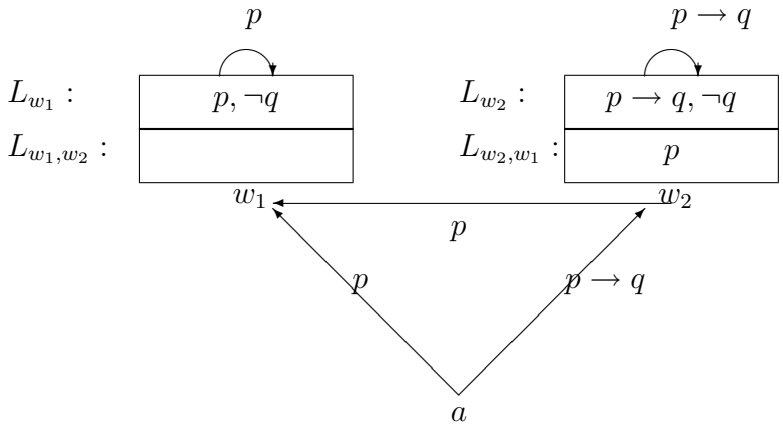
6.4.1. Definition. (Combined Sub-belief)
$L_{i,I}^-(K) \stackrel{\text{def}}{\Longleftrightarrow} \{\varphi \in \mathbf{L_D} : (\exists j \in I)(K \models D_{i,j}\varphi \wedge D_{j,j}\varphi \wedge L_j\varphi)\}$.

The notion of almost safety is obtained from the safety notion by allowing for one more propositional attitude for an informant with respect to the consequences of the

contributed information (clause (ii) below):

6.4.2. DEFINITION. (Almost safety) *For an agent $a \in A_n$ and a DB set K for a, if $C(a, K) = \langle a, I, \psi, \Psi\rangle$ is the induced configuration for a from K, $\{\psi_i\}_{i\in I}$ is said to be* almost safe *for a in K iff for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then, either (i) there exists $i \in I$ such that $L_i^-(K) \not\models \neg\varphi$, or (ii) there exists an $i \in I$ such that $L_{i,I}^-(K)$ is consistent and $L_{i,I}^-(K) \models \varphi$.*

Remarks: The condition (i) gurantees that safety implies almost safety, we show that later. The condition (ii) describes that for any fact φ which is implied by the obtained information is supported by some informant's combined sub-beliefs as well.



Example (a)

6.4.3. EXAMPLE. *A DB set for a, $K = \{L_{w1}(p \wedge \neg q),\ L_{w2}((p \to q) \wedge \neg q),\ D_{w1,w1}p,$ $D_{w2,w2}(p \to q),\ D_{w2,w1}p, L_{a,w1}p,\ L_{a,w2}(p \to q)\}$*

Therefore, $I = \{w1, w2\}$.
$\psi(w1) = p,\ \psi(w2) = p \to q,$
$\Psi(w1) = \{p \wedge \neg q\},$
$\Psi(w2) = \{(p \to q) \wedge \neg q\}.$

So, the induced configuration for a from K is $\langle a, I, \psi, \Psi\rangle$.

$\{\psi_i\}_{i\in I} = \{p, p \to q\}.$

Moreover, from $K \models L_{w1}p \wedge D_{w1,w1}p \wedge D_{w2,w1}p$, and $K \models L_{w_2}(p \to q) \wedge D_{w_2,w_2}(p \to q)$, we have $L_{w2,I}^-(K) = \{p, p \to q\}$ and $L_{w2,I}^-(K)$ is consistent. Evidently, for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then $L_{w2,I}^-(K) \models \varphi$. Thus, $\{\psi_i\}_{i\in I}$ is almost safe for a in K.

The following proposition tells us that the almost safety is really a weaker notion than safety, namely, safety implies almost safety.

6.4.4. PROPOSITION. *Let $a \in A_n$ be an agent, K be a DB set for a, and $C(a, K) = \langle a, I, \psi, \Psi\rangle$ be the induced configuration for a from K.*

If $\{\psi_i\}_{i\in I}$ is safe for a in K, then $\{\psi_i\}_{i\in I}$ is almost safe for a in K.

PROOF. $\{\psi_i\}_{i\in I}$ is safe for a in K, then, by the safety theorem, there exists an $i \in I$ such that $L_i^-(K) \cup \{\psi_i\}_{i\in I}$ is consistent. So, for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then $L_i^-(K) \cup \{\varphi\}$ is consistent, namely, $L_i^-(K) \not\models \neg\varphi$. Therefore, $\{\psi_i\}_{i\in I}$ is almost safe for a in K. □

However, when informants do not rely on each other, almost safety is equal to safety, which is presented by the following proposition.

6.4.5. PROPOSITION. *Let $a \in A_n$ be an agent,K be a DB set for a, and $C(a, K) = \langle a, I, \psi, \Psi\rangle$ be the induced configuration for a from K.*
If K has the following isolated informants property,

for any $i, j \in I$ and any $\varphi \in \mathbf{L_D}$, $i \neq j \Rightarrow K \not\models D_{i,j}\varphi$,

then $\{\psi_i\}_{i\in I}$ is almost safe for a on K iff $\{\psi_i\}_{i\in I}$ is safe for a in K.

PROOF. If K has the isolated informants property.
then, for any $i, j \in I$ and any $\varphi \in \mathbf{L_D}$, $i \neq j \Rightarrow K \not\models D_{i,j}\varphi$. Thus, $L_{i,I}^-(K) = \{\varphi : K \models D_{i,i}\varphi \wedge L_i\varphi\} = \{\varphi : K \models L_{i,i}\varphi\} = L_{i,i}^-(K)$. Note that $L_{i,i}^-(K)$ is always consistent and for any formula χ, $L_{i,i}^-(K) \models \chi \Rightarrow L_i^-(K) \models \chi \Rightarrow L_i^-(K) \not\models \neg\chi$.

$\{\psi_i\}_{i\in I}$ is almost safe for a in K
$\Leftrightarrow$ for any χ, if $\{\psi_i\}_{i\in I} \models \chi$, then either $(\exists i \in I)(L_i^-(K) \not\models \neg\chi)$ or $(\exists i \in I)(L_{i,i}^-(K) \models \chi)$
$\Leftrightarrow$ for any χ, if $\{\psi_i\}_{i\in I} \models \chi$, then $(\exists i \in I)L_i^-(K) \not\models \neg\chi$
$\Leftrightarrow$ $\{\psi_i\}_{i\in I}$ is safe for a in K. □

6.4.6. THEOREM. (Almost safety theorem) *Let $a \in A_n$ be an agent, K be a DB set for a, and $C(a, K) = \langle a, I, \psi, \Psi\rangle$ be the induced configuration for a from K.*
$\{\psi_i\}_{i\in I}$ is almost safe for a in K iff there exists an $i \in I$ such that either $L_i^-(K) \cup \{\psi_i\}_{i\in I}$ is consistent, or $L_{i,I}^-(K)$ is consistent and $L_{i,I}^-(K) \models \{\psi_i\}_{i\in I}$.

PROOF. ($\Rightarrow$) $\{\psi_i\}_{i\in I}$ is almost safe for a in K. Then for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then either there exists $i \in I$ such that $L_i^-(K) \not\models \neg\varphi$ or there exists a $i \in I$ such that $L_{i,I}^-(K)$ is consistent and $L_{i,I}^-(K) \models \varphi$. In particular, $\{\psi_i\}_{i\in I} \models \wedge\{\psi_i\}_{i\in I}$. Thus, either there exists an $i \in I$ such that $L_i^-(K) \not\models \neg \wedge \{\psi_i\}_{i\in I}$, or there exists an $i \in I$ such that $L_{i,I}^-(K)$ consistent and $L_{i,I}^-(K) \models \wedge\{\psi_i\}_{i\in I}$. Therefore, there exists an $i \in I$ such that either $L_i^-(K) \cup \{\wedge\{\psi_i\}_{i\in I}\}$ consistent or $L_{i,I}^-(K)$ consistent and $L_{i,I}^-(K) \models \wedge\{\psi_i\}_{i\in I}$. Moreover, there exists an $i \in I$ such that either $L_i^-(K) \cup \{\psi_i\}_{i\in I}$ consistent or $L_{i,I}^-(K)$ consistent and $L_{i,I}^-(K) \models \{\psi_i\}_{i\in I}$.

($\Leftarrow$) Suppose that there exists an $i \in I$ such that either $L_i^-(K) \cup \{\psi_i\}_{i\in I}$ consistent or

$L^-_{i,I}(K)$ consistent and $L^-_{i,I}(K) \models \{\psi_i\}_{i\in I}$. If $L^-_i(K) \cup \{\psi_i\}_{i\in I}$ is consistent. then, by the safety theorem, $\{\psi_i\}_{i\in I}$ is safe with respect to $\{\Psi_i\}_{i\in I}$. So $\{\psi_i\}_{i\in I}$ is almost safe for a in K. On the other hand, if $L^-_{i,I}(K)$ is consistent and $L^-_{i,I}(K) \models \{\psi_i\}_{i\in I}$, then for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, we have $L^-_{i,I}(K) \models \varphi$ and $L^-_{i,I}(K)$ is consistent. Therefore, $\{\psi_i\}_{i\in I}$ is almost safe for a in K. □

The above theorem offers a general way to test the almost safety for the obtained information. Therefore, we call the following statement *almost-safety test statement.*

(ASTS) there exists an $i \in I$ such that either $L^-_i(K) \cup \{\psi_i\}_{i\in I}$ is consistent, or $L^-_{i,I}(K)$ is consistent and $L^-_{i,I}(K) \models \{\psi_i\}_{i\in I}$.

For a configuration $C = \langle a, I, \psi, \Psi\rangle$, we say that *almost-safety test statement (ASTS) holds in* C iff $(\exists i \in I)(L^-_i(K) \cup \{\psi_i\}_{i\in I}$ is consistent or $(L^-_{i,I}(K)$ is consistent and $L^-_{i,I}(K) \models \{\psi_i\}_{i\in I}))$.

6.5 Almost Safety on Belief Maintenance Operation

In this section we consider the dynamic process of belief revision corresponding to the second stage of the two stage information acquisition process mentioned before. Given a configuration where the receiver has obtained sub-beliefs by hearing statements by thier informants, the receiver will subsequently revise her own belief by incorporation part of these sub-beliefs into her own belief. Clearly she should do so in a safe way; we now have the tools available for formalizing this requirement.

Let $\mathbf{K}$ be a collection of belief sets. As announced before a belief maintenance operation $\triangle : \mathbf{K} \times \mathbf{L_D} \to \mathbf{K}$ is a function assigning a belief set $\triangle(X, \varphi)$ to any belief set $X \in \mathbf{K}$ and each formula φ in $\mathbf{L_D}$.

Our goal is to define an AS operation for the receiver a with respect to the obtained set $\{\psi_i\}_{i\in I}$. Also the revision should lead to the incorporation of the obtained set, since we want to determine under which circumstances it is safe to do so. Recall that $\wedge L^+_{a,I}(\{\psi_i\}_{i\in I})$ and $\wedge\{\psi_i\}_{i\in I}$ denote respectively the compartimentalized belief and the incorporated belief which corresponds the obtained set $\{\psi_i\}_{i\in I}$. In the sequel these two important formulas will be denoted by $\mathbf{cpart}(\psi)$ and $\mathbf{incorp}(\psi)$ respectively.

6.5.1. Definition. (AS operation) *A belief maintenance operation* $\triangle : \mathbf{K} \times L_D \to \mathbf{K}$ *is said to be an almost safety one for agent* $a \in A_n$ *with respect to* $\{\psi_i\}_{i\in I}$*, iff for any DB set* $K \in \mathbf{K}$ *for a such that* $L^-_a(K) \models \mathbf{cpart}(\psi)$ *and* $L^-_a(K) \not\models \mathbf{incorp}(\psi)$*, it will be the case that* $\triangle(L^-_a(K), \mathbf{cpart}(\psi)) \models \mathbf{incorp}(\psi)$ *only when* $\{\psi_i\}_{i\in I}$ *is almost safe for a in* K*.*

Remarks: (i) We define almost safety for a belief maintenance operation in terms of the general almost-safety notion.

(ii) We consider only the case where the knowledge state K is a DB set for a since otherwise the induced configuration does not exist, and consequently the concept of almost safety does not make sense.
(iii) $L_a^-(K) \models \mathbf{cpart}(\psi)$ means that the receiver a has full knowledge about his compartimentalized information $\wedge L_{a,I}^+(\{\psi_i\}_{i\in I})$.
(iv) $L_a^-(K) \not\models \mathbf{incorp}(\psi)$ and $\triangle(L_a^-(K),\mathbf{cpart}(\psi)) \models \mathbf{incorp}(\psi)$ together means that we consider only the case where the receiver a really assimilates the obtained information.

In other words, agent a originally does not fully believe the fact $\wedge\{\psi_i\}_{i\in I}$, but by invoking the operation, she fully believes this fact. The format for our belief revision operator therefore further specializes to:

$$\triangle(L_a^-(K),\mathbf{cpart}(\psi)) = \begin{cases} L_a^-(K)\dot{+}\mathbf{incorp}(\psi) & \text{if } \varphi_{i_1} \in L_a^-(K) \text{ or ... or } \varphi_{i_n} \in L_a^-(K) \\ L_a^-(K) & \textbf{otherwise} \end{cases}$$

Recall that the revision operator $\dot{+}$, in this type belief maintenance operation, satisfies the success postulate (d), i.e., $\varphi \in K\dot{+}\varphi$. We can characterize almost safety:

6.5.2. Theorem. (AS operation theorem) *Consider a belief maintenance operation $\triangle$ of the form:*

$$\triangle(L_a^-(K),\mathbf{cpart}(\psi)) = \begin{cases} L_a^-(K)\dot{+}\mathbf{incorp}(\psi) & \textit{if } \varphi_{i_1} \in L_a^-(K) \textit{ or ... or } \varphi_{i_n} \in L_a^-(K) \\ L_a^-(K) & \textbf{otherwise} \end{cases}$$

Operation $\triangle$ is an AS operation for a with respect to $\{\psi_i\}_{i\in I}$ iff every assumption in the sequence $\varphi_{i_1} \in L_a^-(K)$ or ... or $\varphi_{i_n} \in L_a^-(K)$ entails that the almost-safety test holds in $C(a, K)$.

PROOF. ($\Rightarrow$)
$\triangle$ is an AS operation for a with respect to $\{\psi_i\}_{i\in I}$
$\Rightarrow$ [for any DB set K for a such that $L_a^-(K) \models \mathbf{cpart}(\psi)$ and $L_a^-(K) \not\models \mathbf{incorp}(\psi)$, $(\triangle(K,\mathbf{cpart}(\psi)) \models \mathbf{incorp}(\psi) \Rightarrow \{\psi_i\}_{i\in I}$ is almost safe for a in $K)$] (Definition of AS operation)
$\Rightarrow$[for any DB set K for a such that $L_a^-(K) \models \mathbf{cpart}(\psi)$ and $L_a^-(K) \not\models \mathbf{incorp}(\psi)$, $(\varphi_{i_1} \in K$ or $\varphi_{i_2} \in K$ or ... or $\varphi_{i_n} \in L_a^-(K) \Rightarrow \mathbf{incorp}(\psi)$ is almost safe for a in $K)$] (Definition of $\triangle$ and $L_a^-(K) \not\models \mathbf{incorp}(\psi)$ and the success postulate)
$\Rightarrow$ [for any DB set K for a such that $L_a^-(K) \models \mathbf{cpart}(\psi)$ and $L_a^-(K) \not\models \mathbf{incorp}(\psi)$, $(\varphi_{i_1} \in K$ or $\varphi_{i_2} \in K$ or ... or $\varphi_{i_n} \in L_a^-(K) \Rightarrow$ the almost-safety test statement holds)] (Almost safety theorem)

($\Leftarrow$) It is straightforward from the definition of AS operation. □

Evidently, our goal is to define an AS operation for general cases. There remain however complications. For example it is not possible to check that a set of formulas K is consistent by testing whether particular formulas belong to K or not. Therefore we need some further assumptions. We only consider DB sets K for a for which

the combined sub-belief sets $L^-_{i,I}(K)$ are consistent. Another additional condition is that we only consider DB set K for an agent a for which knowledge and belief coincide: $K \models L_a\varphi \rightarrow \varphi$ for any φ. We call such an agent *a a skeptic agent in* K. An operation which is AS under the above two additional assumptions will be called a *restricted AS operation.*

6.5.3. DEFINITION. (Restricted AS operation) *A belief maintenance operation* $\triangle : \mathbf{K} \times L_D \rightarrow \mathbf{K}$ *is said to be a restricted almost safe one for agent* $a \in A_n$ *with respect to* $\{\psi_i\}_{i\in I}$*, iff for any DB set* $K \in \mathbf{K}$ *for a such that (i)* $L^-_a(K) \models$**cpart**(ψ)*, (ii)* $L^-_a(K) \not\models$**incorp**(ψ)*, (iii) agent a is a skeptic agent in* K*, and (iv) any combined sub-belief set from* K *is consistent, it holds that* $\triangle(L^-_a(K),$**cpart**$(\psi)) \models$**incorp**(ψ) *only when* $\{\psi_i\}_{i\in I}$ *is almost safe for a in* K*.*

6.5.4. THEOREM. (Restricted AS operation theorem) *Suppose that a belief maintenance operation* $\triangle$ *is a type 5 operation like:*

$$\triangle(L^-_a(K),\mathbf{cpart}(\psi)) = \begin{cases} L^-_a(K)\dot{+}\{\psi_i\}_{i\in I} & \text{if } \varphi_{i_1} \in L^-_a(K) \text{ or } \ldots \text{ or } \varphi_{i_n} \in L^-_a(K) \\ L^-_a(K) & \textbf{otherwise} \end{cases}$$

$\triangle$ *is a restricted AS operation for a with respect to* $\{\psi_i\}_{i\in I}$ *iff for every belief state* K *satisfying the conditions (i), (ii), (iii) and (iv) above, every assumption in the sequence* $\varphi_{i_1} \in L^-_a(K)$ *or ... or* $\varphi_{i_n} \in L^-_a(K)$ *entails that the almost-safety test holds in* $C(a, K)$*.*

PROOF. Similar to the proof for AS theorem. □

After these preparations we are finally ready to define a restricted AS operation for the agent a with respect to the obtained information $\{\psi_i\}_{i\in I}$ where $I = \{i_1, \ldots i_k\}$. The defined operation considers two kinds of typical situations. The first situation is that each informant fully relies both on other informants and on herself about what they say, i.e., $\bigwedge_{j=1}^k \bigwedge_{j'=1}^k D_{i_j,i_{j'}}\psi(i_{j'})$. In this situation, every informant plays a role of an "expert" on the information she offers as defined in the last chapter. Note that for each informant i_l, the above condition can be reduced to the condition $\bigwedge_{j=1}^k D_{i_l,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^k D_{i_j,i_j}\psi(i_j)$.

The second situation is one which already supports a stronger notion of safety, meaning that some informant i_l considers the obtained set consistent with her beliefs, i.e., $\neg L_{w_{i_l}} \neg \bigwedge_{j=1}^k \psi(i_j)$. Since according to the honesty condition, each informant i_l already believes what she offers, the above condition can be weakened to the less restrictive condition $\neg L_{w_{i_l}} \neg \bigwedge_{j=1,j\neq l}^k \psi(i_j)$. Formally we define our operation as a type 5 operation as follows:

The Definition of Operation $\triangle_{ras1}$ **(for Agent** a**):**

(A1) $\bigwedge_{j=1}^k D_{i_1,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^k D_{i_j,i_j}\psi(i_j) \Rightarrow L^-_a(K) \triangle_{ras1} \mathbf{cpart}(\psi) = L^-_a(K)\dot{+}\mathbf{incorp}(\psi)$.

(A2) $\bigwedge_{j=1}^{k} D_{i_2,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j) \Rightarrow L_a^-(K) \triangle_{ras1} \mathbf{cpart}(\psi) = L_a^-(K) \dot{+} \mathbf{incorp}(\psi)$.

.....

(Ak) $\bigwedge_{j=1}^{k} D_{i_k,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j) \Rightarrow L_a^-(K) \triangle_{ras1} \mathbf{cpart}(\psi) = L_a^-(K) \dot{+} \mathbf{incorp}(\psi)$.

(B1) $\neg L_{i_1} \neg \bigwedge_{j=2}^{k} \psi(i_j) \Rightarrow L_a^-(K) \triangle_{ras1} \mathbf{cpart}(\psi) = L_a^-(K) \dot{+} \mathbf{incorp}(\psi)$.

(B2) $\neg L_{i_2} \neg \bigwedge_{j=1,j\neq 2}^{k} \psi(i_j) \Rightarrow L_a^-(K) \triangle_{ras1} \mathbf{cpart}(\psi) = L_a^-(K) \dot{+} \mathbf{incorp}(\psi)$.

.....

(Bk) $\neg L_{i_k} \neg \bigwedge_{j=1}^{k-1} \psi(i_j) \Rightarrow L_a^-(K) \triangle_{ras1} \mathbf{cpart}(\psi) = L_a^-(K) \dot{+} \mathbf{incorp}(\psi)$.

For the above operation, the cases (A1)-(Ak) are representative for the original problem as posed by Schoenmakers, since we need no further information about source agents' beliefs other than the general information about the rely-on relations among agents. The cases (B1)-(Bk) deal with the situation where agent a may have previously collected some information about the source agents' beliefs and the obtained information is already safe. Although these situations are not representative for our problem, handling those situation is necessary for obtaining a more general operation.

6.5.5. THEOREM. *The operation $\triangle_{ras1}$ is a restricted AS operation for agent a with respect to $\{\psi_i\}_{i\in I}$.*

PROOF. Let $A(l) = \bigwedge_{j=1}^{k} D_{i_l,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j)$, where $l \in \{1, 2, ..., k\}$; and $B(l) = \neg L_{i_l} \neg \bigwedge_{j=1,j\neq l}^{k}$, where $l \in \{1, 2, ..., k\}$;

We have to show that ($A(1) \in L_a^-(K)$ or $A(2) \in L_a^-(K)$ or ... or $A(k) \in L_a^-(K)$ or $B(1) \in L_a^-(K)$ or ... or $B(k) \in L_a^-(K)$) implies that (ASTS) holds in $C(a, K)$.

It is sufficient to show that (1) $A(l) \in L_a^-(K) \Rightarrow L_{i_l,I}^-(K) \models \{\psi_i\}_{i\in I}$, and
(2) $B(l) \in L_a^-(K) \Rightarrow L_{i_l}^-(K) \cup \{\psi_i\}_{i\in I}$ is consistent, for $1 \leq l \leq k$.

Case (1)
$A(l) \in L_a^-(K)$
$\Rightarrow \bigwedge_{j=1}^{k} D_{i_l,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j) \in L_a^-(K)$ (By definition of $A(l)$)
$\Rightarrow K \models L_a(\bigwedge_{j=1}^{k} D_{i_l,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j))$ (By definition of $L_a^-(K)$)
$\Rightarrow K \models \bigwedge_{j=1}^{k} D_{i_l,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j)$ (Since a is a skeptic agent)
$\Rightarrow K \models \bigwedge_{j=1}^{k} (D_{i_l,i_j}\psi(i_j) \wedge D_{i_j,i_j}\psi(i_j) \wedge L_{i_j}\psi(i_j))$ (By the honesty condition)
$\Rightarrow \{\psi_i\}_{i\in I} \subseteq L_{i_l,I}^-(K)$ (By definition of $L_{i_l,I}^-(K)$)
$\Rightarrow L_{i_l,I}^-(K) \models \{\psi_i\}_{i\in I}$.

Case (2)
$B(l) \in L_a^-(K)$
$\Rightarrow \neg L_{i_l} \neg \bigwedge_{j=1,j\neq l}^{k} \psi(i_j) \in L_a^-(K)$ (By definition of $B(l)$)
$\Rightarrow K \models L_a(\neg L_{i_l} \neg \bigwedge_{j=1,j\neq l}^{k} \psi(i_j))$ (By definition of $L_a^-(K)$)
$\Rightarrow K \models \neg L_{i_l} \neg \bigwedge_{j=1,j\neq l}^{k} \psi(i_j)$ (Since a is a skeptic agent)
$\Rightarrow K \not\models L_{i_l} \neg \bigwedge_{j=1,j\neq l}^{k} \psi(i_j)$ (Since K is consistent)
$\Rightarrow L_{i_l}^-(K) \not\models \neg \bigwedge_{j=1,j\neq l}^{k} \psi(i_j)$ (By definition of $L_{i_l}^-(K)$)
$\Rightarrow L_{i_l}^-(K) \not\vdash \neg \bigwedge_{j=1,j\neq l}^{k} \psi(i_j)$ (By the soundness)
$\Rightarrow L_{i_l}^-(K) \cup \{\bigwedge_{j=1,j\neq l}^{k} \psi(i_j)\}$ is consistent. (Meta reasoning)
$\Rightarrow L_{i_l}^-(K) \cup \{\psi_i\}_{i\in I}$ is consistent. (By honesty) □

Using the definition of the operation $\triangle_{ras1}$ and the above theorem, it becomes a straightforward application to construct a restricted AS operation for the judge; just consider the special case where $I = \{w_1, w_2\}$ and $\{\psi_i\}_{i\in I} = \{\psi(w_1) = p, \psi(w_2) = p \rightarrow q\}$, i.e., the agent w_1 offers information p, and agent w_2 offers information $p \rightarrow q$.

The Definition of Operation $\triangle_{jp}$ (for Agent a):
(A1) $D_{w_1,w_2}(p \rightarrow q) \wedge D_{w_1,w_1} p \wedge D_{w_2,w_2}(p \rightarrow q) \Rightarrow L_a^-(K) \triangle_{jp} L_{a,w_1} p \wedge L_{a,w_2}(p \rightarrow q) = L_a^-(K) \dot{+} p \wedge (p \rightarrow q)$.

(A2) $D_{w_2,w_1} p \wedge D_{w_1,w_1} p \wedge D_{w_2,w_2}(p \rightarrow q) \Rightarrow L_a^-(K) \triangle_{jp} L_{a,w_1} p \wedge L_{a,w_2}(p \rightarrow q) = L_a^-(K) \dot{+} p \wedge (p \rightarrow q)$.

(B1) $\neg L_{w_1} \neg q \Rightarrow L_a^-(K) \triangle_{jp} L_{a,w_1} p \wedge L_{a,w_2}(p \rightarrow q) = L_a^-(K) \dot{+} p \wedge (p \rightarrow q)$.

(B2) $\neg L_{w_2} \neg p \Rightarrow L_a^-(K) \triangle_{jp} L_{a,w_1} p \wedge L_{a,w_2}(p \rightarrow q) = L_a^-(K) \dot{+} p \wedge (p \rightarrow q)$.

There remains the task of presenting this rather intricate solution in some more conceptual way. In order to explain our solution to someone who understands the original puzzle but is not able to grasp the full power of the logic machinery called into action, we can present a new sequel to the Judge puzzle story which leads to an unexpected solution. Assuming that the judge drew his conclusion based on our restricted AS operation, we discover that the unacceptability of the state of affairs as indicated by the original story only is represents a temporary stage in the process of exchanging information and incorporation of beliefs. The continuation of the story (the part which Schoenmakers did not include in his paper) goes as follows:

> *When the judge was told that p was true by the witness w*1 *and learned that the implication p → q was true from witness w*2*, she had to figure out whether these assertions could be accepted together. Now the judge had*

good reasons for not asking the witnesses for more information about their knowledge, since she could base her decision already on her knowledge of the rely-on relation. She knew that witness $w1$ was the only authority concerning the statement p, and that witness $w2$ was the only authority concerning the conditional $p \rightarrow q$. Moreover, this information was common knowledge among both witnesses and herself. Therefore, she could safely conclude that q was true, and consequently she ordered the verdict to be executed. When they learned about this execution both witnesses $w1$ and $w2$ came forward and protested against the verdict, claiming that q was false. The judge patiently informed witness $w1$ about the witness $w2's$ belief that $p \rightarrow q$ was true. Because the witness $w1$ accepted that $w2$ was the authority on the implication $p \rightarrow q$, $w1$ accepted this assertion, and had to agree with the judge. She also told witness $w2$ about $w1$'s belief, that p was true, and consequently witness $w2$ also had to agree with her verdict, since $w2$ accepted that the $w1$ was the authority about p. In the end everybody was satisfied.

6.6 Conclusions

We have formalized the problem of information acquisition in a multi agent environment. The danger of accepting information from several agents as illustrated in the judge puzzle is an inherent consequence of disagreement among the informants; there exists no absolute safe set of obtained information other than trivial sets, and safe or strongly safe sets are defined only relative the full believe state which in general is unknown to the receiver.

Formalizing this problem in a belief dependence framework does not offer an easy way out; however, by assuming the initial role-knowledge assumption, honesty, skepticism for the judge and a few consistency conditions, and by considering a highly specialized belief maintenance operation a restricted almost safe solution for the judge puzzle has been obtained. This solution has moreover the nice property that it is computable.

Notwithstanding its complexity, our solution has some interesting features: it is based on a general theory supported by psychological evidence, and the tools used for the solution were not developed for the purpose of solving the Judge puzzle. We consider it highly unlikely that there exist "cleaner" solutions to this problem (aside of simply denying it to be a problem).

For designers of intelligent database systems and expert systems our results suggest the following guideline: When combining expertise from different expert sources, ensure that the contributing agents involved recognize each other to be the expert on their respective contributions. If the situation should ever arise that some contributing agent starts complaining about the knowledge stored in the system, the designer, by following our guideline, has ensured that during the subsequent debate she won't be forced to redesign the knowledge base; instead the complaining informants will learn something they didn't know before.

Chapter 7

Conclusions and Further Work

7.1 Concluding Remarks

In this thesis, a formalism for logics of belief dependence has been proposed. Several axiom systems for those logics have been suggested. Their corresponding semantic models have been introduced. The soundness, completeness, complexity, and decidability problems for some of the proposed logics have been studied. Furthermore, several applications of the logics of belief dependence have been discussed. One of the applications involves the use of belief dependence logics for capturing an algorithm to guide rational agents in their belief dynamics. Another application involves belief dependence for a plausible solution to Schoenmakers problem. The results show that the proposed logics indeed is a promising and expressive tool to study the problems concerning information communication and belief dependence in a multi-agent environment.

7.2 Further Work

There is still a lot of further work to do in the studies of logics of belief dependence. Here are some possible extensions:

7.2.1 More General Semantic Models

We have proposed several semantic models such as, D-model and Lij-model, for the logic of belief dependence. Those models are suitable only for their own special application cases. For instance, the logics which are based on D-models all obey the honesty axiom $L_{i,j}\varphi \rightarrow L_j\varphi$, which is not a very nice property if we want to formalize a system in which the agents are not always honest. On the other hands, because of the necessitation rule (NECLij), the logics based on Lij-models make the systems suffering from the problem of logical omniscience. A natural attempt is to combine the D-model approach with Lij-model approach to achieve a new model. We

can modify D-models by extending the accessibility relations function $\mathcal{L}$ in D-models into their counterpart in Lij-models. a possible definition reads:

7.2.1. Definition. (Belief dependence LijD-model) *A belief dependence LijD-model is a tuple $M = \langle W, \mathcal{L}, \mathcal{D}, V\rangle$ where W is a set of possible worlds, V is a truth assignment function, and $\mathcal{L} : \mathbf{A_n} \times \mathbf{A_n} \to \mathcal{P}(W \times W)$, consists of $n \times n$ binary accessibility relations on W,$\mathcal{D} : \mathbf{A_n} \times \mathbf{A_n} \times W \to \mathcal{P}(\mathbf{L_{LijD}})$, a relation describing the dependencies as before.*

The truth condition $\models$ is defined inductively as follows:

$M, w \models p$, where p is a primitive proposition, iff $w \in V(p)$,
$M, w \models \neg\varphi$ iff $M, w \not\models \varphi$,
$M, w \models \varphi_1 \wedge \varphi_2$ iff $M, w \models \varphi_1$ and $M, w \models \varphi_2$,
$M, w \models L_{i,j}\varphi$ iff $M, t \models \varphi$ for all t such $\langle w, t\rangle \in \mathcal{L}(i, j)$
$M, w \models D_{i,j}\varphi$ iff $\varphi \in \mathcal{D}(i, j, w)$.

Similaring to those in Lij-models, in LijD-models, we still consider $L_{i,i}\varphi$ as its standard epistemic interpretation, namely, $L_i\varphi$. The minimal logic system, called **LijD** is the system with axioms (BA), (KLij), inference rules (MP) and (NECLij), and the definition (Ldf).

7.2.2. Theorem. **LijD** *belief dependence systems are sound and complete in the class of LijD-models.*

PROOF. Combining the proof for the system **LD** and the proof for the system **Lij**, we also can show that soundness and completeness of the system **LijD**. The details are omitted. □

Although in the systems based on LijD-model the necessitation rule (NECLij) still holds, we can avoid the problem of logical omniscience by introducing an explicit belief operator B_i as follows:

$$B_i\varphi \stackrel{\text{def}}{\Longleftrightarrow} L_i\varphi \wedge D_{i,i}\varphi.$$

7.2.1. Claim. *In* **LijD** *belief dependence systems, explicit beliefs are not closed under valid implication.*

Naturally, there exists still a lot of further work to do on this LijD-model. For instances, more extensions of the minimal logic system, and their soundness, completeness, and decidability problems, are open for further research.

7.2.2 Other Complexity Problems

In this thesis, we only study the complexity problems for some logical systems. However, there exist a lot of different logical systems of belief dependence. Although one cannot exhaust all of these logics systems, the complexity results for some interesting system are still interesting.

We state the following conjectures:

i) The satisfiability problem of $\mathbf{L5^-}$+**D4** system with one agent is NP-complete.

ii) The satisfiability problem of $\mathbf{L5^-}$+**D4** system with more than one agent is PSPACE-complete.

iii) The satisfiability problem of $\mathbf{Lij5^-}$+**D** system with more than one agent is PSPACE-complete.
(Note that by Halpern and Moses' results, we already know that the satisfiability problem of $\mathbf{Lij5^-}$+**D** system with one agent is NP-complete.)

7.2.3 Alternative Almost Safety Belief Maintenance Operations

In the chapter 6, we offer a definition of an AS operation, in which the agent assimilate the information only when the information is almost safe. A stronger definition for an AS operation is one where the agent assimilates the information *if and only if* the information is almost safe. Formally, we have the following definition:

7.2.3. DEFINITION. (Complete AS Operation) *A belief maintenance operation* $\triangle : \mathbf{K} \times L_D \rightarrow \mathbf{K}$ *is said to be a complete AS operation for agent* $a \in A_n$ *with respect to* $\{\psi_i\}_{i\in I}$*, iff for any DB set* $K \in \mathbf{K}$ *for a such that* $L_a^-(K) \models \mathbf{cpart}(\psi)$ *and* $L_a^-(K) \not\models \mathbf{incorp}(\psi)$*, it will be the case that* $\triangle(L_a^-(K),\mathbf{cpart}(\psi)) \models \mathbf{incorp}(\psi)$ *if and only iff* $\{\psi_i\}_{i\in I}$ *is almost safe for a in* K*.*

A complete AS operation is more useful, since it requires that the agent assimilates as much information as possible. Obtaining more results for a complete AS operation remains an interesting challenge.

Part II

Action Logics for Agents with Bounded Rationality

Chapter 8

Introduction

8.1 Motivation

This chapter provides an introduction to the study of action logics for agents with bounded rationality in order to develop a formal language for social science theories, in particular for theories of organization and management.

Present day theories in social sciences are either expressed in natural language or in a small subset of equational mathematical theories. They lack a formal foundation that would allow to check their consistency in a rigorous fashion, or to disambiguate natural language statements. As a consequence, these theories have acquired a reputation for "softness" – a soft way of saying that their logical properties are somewhat dubious. Reformulating them in a formal language with known properties would facilitate the tasks of consistency checking or disambiguation. Also, it would prepare the ground for other tasks, for instance the investigation of deductive closure properties.

We focus on action logic, because actions (of individual or collective agents) are key to the understanding of social phenomena. In fact, most social scientist agree that adequate theories of social relations must be action theories first [Blumer 1969, Giddens 1979, Luhmann 1982, Parsons 1937, Schutz 1967]. Yet actions generate change, and change is notoriously hard to grasp in the extensional context of first order languages [Gamut 1990]. This justifies our attempt to develop a new logic, rather than taking First Order Logic off the shelf. This new logic is called ALX (the x'th action logic).

Action logics are not new in formal AI. There have been a variety of attempts to put actions into a logical framework starting with McCarthy and Hayes's seminal paper [McCarthy&Hayes 1969], see in particular [Cohen&Levesque 1987, Jackson 1989, Cohen&Levesque 1990, Ginsberg&Smith 1987, Rao&Georgeff 1991, Winslett 1988]. However, present day action logics are usually developed for the (hypothetical) use by intelligent robots[Cohen&Levesque 1990, Ginsberg&Smith 1987, Rao&Georgeff 1991] or as a description language of program behavior [Harel 1984]. Our effort is motivated by a different concern. The difference in motivation leads to a new approach to action

logic, combining ideas from various strands of thought, notably H.A. Simon's notion of *bounded rationality*, G. H.von Wright's approach to *preferences*, Kripke's *possible world semantics*, J. Hintikka's approach to *knowledge and beliefs*, Pratt's *dynamic logic* in combination with Stalnaker's notion of *minimal change*, and more recent ideas from *belief revision* and *update semantics* [Grahne 1991, Hirofumi&Mendelzon 1991].

8.2 General Considerations

Herbert A. Simon's conceptualization of *bounded rationality* [Simon 1955] serves as a point of departure. His approach is intuitively appealing, and had great impact on the postwar social sciences. Simon wanted to overcome the omniscience claims of the traditional conceptualizations of rational action. He assumed (1) an agent with (2) a set of behavior alternatives, (3) a set of future states of affairs (each such state being the outcome of a choice among the behavior alternatives), and (4) a preference order over future states of affairs. The omniscient agent, endowed with "perfect rationality", would know all behavior alternatives and the exact outcome of each alternative; the agent would also have a complete preference ordering for those outcomes. An agent with bounded rationality, in contrast, would not know all alternatives, nor would she know the exact outcome of each; also, the agent would lack a complete preference ordering for those outcomes.

Kripke's *possible world semantics* provides a natural setting for Simon's conceptualization. We assume a set of possible worlds with various relations defined over this set (we may also call those possible worlds *states*) . One can see a behavior alternative as a mapping from states to states, so each behavior alternative constitutes an accessibility relation. An accessibility relation, in turn, can be interpreted as an opportunity for action, i.e., as an opportunity for changing the world by moving from a given state to another state. Accessibility relations are expressed by indexed one-place modal operators, as in dynamic logic [Harel 1984]. For example, the formula $\langle a \rangle \phi$ would express the fact that the agent has an action a at her disposal such that effecting a in the present situation would result in the situation denoted by proposition ϕ.

The perfectly rational agent would have a complete description of her actual state, a complete knowledge of all accessibility relations, and a complete preference ordering over states. Agents with bounded rationality are less well informed. They have an incomplete description of their actual state (we call those descriptions *situations*), incomplete knowledge of the accessibility relations, and an incomplete preference ordering over situations.

Situations are represented as sets of states and expressed by propositions. Propositions, in turn, denote the set of states where they obtain. So, the more specific an agent's knowledge about a situation, the more detailed the propositional description of that situation would be. The limit case, a complete description, would uniquely identify one state. Less specific descriptions would lack that uniqueness, identifying the set of those states where the description would hold (but remaining uncom-

Figure 8.1: Simon's Bounded Rationality

mitted about other "aspects" not covered by the description). This is a standard approach to representing incomplete information, used in denotational semantics [Scott 1970, Scott 1982] and epistemic logic [Halpern&Moses 1992]. [1]

Preferences – not goals – provide the basic rationale for rational action in ALX. Following von Wright [von Wright 1963], a preference statement is understood as a statement about situations. For example, the statements that "I prefer oranges to apples" is interpreted as the fact that "I prefer the states in which I have an orange to the states in which I have an apple." Following v. Wright again, we assume that an agent who says to prefer oranges to apples should prefer a situation where she has an orange but *no* apple to a situation where she has an apple but *no* orange. We call this principle *conjunction expansion principle* and restrict attention to preference statements that obey it. Preferences are expressed via two-place modal operators; if the agent prefers the proposition ϕ to the proposition ψ, we write $\phi \boldsymbol{P} \psi$.

Normally, the meaning of a preference statement is context-dependent, even if this is not made explicit. An agent may say to prefer an apple to an orange – and actually mean it – but she may prefer an orange to an apple later – perhaps because then she already had an apple. To capture this context dependency, we borrow the notion of minimal change from Stalnaker's approach to conditionals [Stalnaker 1968]. The idea is to apply the conjunction expansion principle only to situations that are minimally different from the agent's present situation – just as different as they really need to be in order to make the propositions true about which preferences are expressed. We introduce a binary function, *cw*, to the semantics that determines the set of "closest" states relative to a given state, such that the new states fulfil some specified conditions, but resembles the old state as much as possible in all other respects. For situations (sets of states) we apply *cw* to each element of the situation separately. This allows us to avoid some technical problems arising in conditional logic [Lewis 1973, Nute 1986, Stalnaker 1968].

ALX provides a complete syntactic characterization of preferences, so one can derive new preference statements from old ones by using its machinery. Closing the set of preference statements under the rules of inferences yields a preference order that serves as the basis for deriving goals. *Goals*, in turn, can be defined as preferred (in some sense) and accessible (or at least not believed to be inaccessible) situations; once an agent has inferred her goals, she is supposed to plan the chain(s) of actions that would get her to a goal. Note that goals need not be unique. Also, the goal set need not be closed under logical implication, so agents need not treat undesired consequences of desired outcomes (e.g., tooth ache as a consequence of having one's teeth repaired) as goals (as opposed to action logics that use the concept of "goal" as the primitive notion of rational guidance).

[1] Framing bounded rationality in terms of possible worlds semantics reveals a fine point usually ignored: one can see that omniscience – the limiting case – is contingent upon the choice of the language. Full rationality in an absolute sense would require a language isormorphic to the universe "out there", but such a language is not available. Any formal theory about full rationality has to make simplifying assumptions about the world, but those assumption, by their simplifying nature, violate the ontology of full rationality in some sense or another.

As we can see, our approach heavely depends on the notion of the minimal change and conditional logics. Therefore, in the next a few sections, let us examine the notions of minimal change, conditional, update, and their relationship in details.

8.3 Conditional and Update

8.3.1 Counterfactuals and Conditional Logic

A counterfactual is a statement such as, "if p would be the case, then q", where the premise p is either known or expected to be false. In actual practice, a counterfactual is generally represented as a subjunctive conditional. A counterfactual conditional is written as "$p \rightsquigarrow q$" to denote "if p, then q", to distinguish it from the material conditional "$p \rightarrow q$".

Counterfactuals are not truth-functional, because they can only be evaluated relative to (i) some theory of what the world is like, and (ii) some notions about what the world might be like if certain things were to change. Stalnaker [Stalnaker 1968] was the first to propose a possible world semantics for the counterfactual. In his approach, the conditional $\psi \rightsquigarrow \phi$ is true just in case ϕ is true at the world most like the actual world at which ψ is true. In order to determine the truth of a conditional counterfactual, we add the premise of the conditional to our set of beliefs and modify this set as little as possible in order to accommodate the new premise. Finally, we consider whether the consequent of the conditional would be true, were the revised set of beliefs all true. As a consequence, the corresponding formal semantics must specify the meaning of "minimal change". Stalnaker assumes that there is always a unique possible world at which the premise is true and which is more like the actual world than any other world at which the premise is true, provided the premise of a conditional is logically possible. This assumption is called *Stalnaker's Uniqueness Assumption.*

Stalnaker's semantic model is an ordered quadruple $\langle W, R, s, [\ ] \rangle$ where W is a set of possible worlds, R is a binary reflexive accessibility relation on W, s is a world selection function which assigns to a sentence ψ and a world w in W a world $s(\psi, w)$, called the ψ-world closest to w, and $[\ ]$ is a function which assigns to each sentence ψ a subset $[\psi]$ of W. Moreover, the models satisfy the following constraints:

(S1) $s(\psi, w) \in [\psi]$.
(S2) $\langle w, s(\psi, w) \rangle \in R$.
(S3) if $s(\psi, w)$ is not defined then for all $w' \in I$ such that $\langle w, w' \rangle \in R, w' \notin [\psi]$.
(S4) if $w \in [\psi]$ then $s(\psi, w) = w$.
(S5) if $s(\psi, w) \in [\phi]$ and $s(\phi, w) \in [\psi]$, then $s(\psi, w) = s(\phi, w)$.
(S6) $w \in [\psi \rightsquigarrow \phi]$ if and only if $s(\psi, w) \in [\phi]$ or $s(\psi, w)$ is undefined.

The conditional logic determined by Stalnaker's model theory is the smallest conditional logic which is closed under the two inference rules

(RCEC) $\vdash \psi \leftrightarrow \phi \Rightarrow \vdash (\chi \rightsquigarrow \psi) \leftrightarrow (\chi \rightsquigarrow \phi)$.
(RCK) $\vdash (\psi_1 \wedge ... \wedge \psi_n) \rightarrow \phi \Rightarrow \vdash ((\chi \rightsquigarrow \psi_1) \wedge ... \wedge (\chi \rightsquigarrow \psi_n)) \rightarrow (\chi \rightsquigarrow \phi),\ n \geq 0$.

and which contains all substitution instances of the following axioms together with *modus ponens* and the set of tautologies.

(ID) $\psi \rightsquigarrow \psi$.
(MPC) $(\psi \rightsquigarrow \phi) \rightarrow (\psi \rightarrow \phi)$.
(MOD) $(\neg\psi \rightsquigarrow \psi) \rightarrow (\phi \rightsquigarrow \psi)$.
(CSO) $[(\psi \rightsquigarrow \phi) \wedge (\phi \rightsquigarrow \psi)] \rightarrow [(\psi \rightsquigarrow \chi) \leftrightarrow (\phi \rightsquigarrow \chi)]$.
(CV) $[(\psi \rightsquigarrow \phi) \wedge \neg(\psi \rightsquigarrow \neg\chi)] \rightarrow [(\psi \wedge \chi) \rightsquigarrow \phi]$.
(CEM) $(\psi \rightsquigarrow \phi) \vee (\psi \rightsquigarrow \neg\phi)$.

This system is called **C2**. Lewis disagrees with Stalnaker's Uniqueness Assumption, and suggests to drop (CEM), the Conditional Excluded Middle [2]. In [Lewis 1973], Lewis proposes several semantic models which can avoid the conditional excluded middle. One of his systems is as follows:

The minimal change model M is an ordered triples $\langle W, cw, V\rangle$, where W and V are, as before, a set of possible worlds and a valuation function, and $cw : W \times \mathcal{P}(W) \hookrightarrow \mathcal{P}(W)$ is a partial function which assigns to each world w in W and a subset of W a subset of W, which is called *class selection function*, or *closest world function* because it assumes that there is at least one closest ψ-world for our selection function to pick out if ψ is possible at w. Moreover, the class selection function cw satisfies the following constraints.

(CS1) $cw(w, [\![\phi]\!]_M) \subseteq [\![\phi]\!]_M$.
(CS2) $w \in [\![\phi]\!]_M \Rightarrow cw(w, [\![\phi]\!]_M) = \{w\}$.
(CS3) $cw(w, [\![\phi]\!]_M) = \emptyset \Rightarrow cw(w, [\![\psi]\!]_M) \cap [\![\phi]\!]_M = \emptyset$.
(CS4) $cw(w, [\![\phi]\!]_M) \subseteq [\![\psi]\!]_M$ and $cw(w, [\![\psi]\!]_M) \subseteq [\![\phi]\!]_M \Rightarrow cw(w, [\![\phi]\!]_M) = cw(w, [\![\psi]\!]_M)$.
(CS5) $cw(w, [\![\phi]\!]_M) \cap [\![\psi]\!]_M \neq \emptyset \Rightarrow cw(w, [\![\phi \wedge \psi]\!]_M) \subseteq cw(w, [\![\phi]\!]_M)$.

where $[\![\phi]\!]_M \overset{\text{def}}{\Longleftrightarrow} \{w \in W : M, w \models \phi\}$.

The truth condition for the conditional is defined as follows:
$M, w \models \psi \rightsquigarrow \phi$ if and only if $cw(w, [\![\psi]\!]_M) \subseteq [\![\phi]\!]_M$.

The above model determines a smallest conditional logic which is closed under the same rules as those listed for **C2** except that we replace (CEM) with (CS):

[2] There are several counterexamples against (CEM). One of them is as follows: We believe intuitively that there is a distinction between the meaning of the counterfactual conditional $\psi \rightsquigarrow \phi$ and that of "might be" conditional $\psi \diamond\!\!\rightarrow \phi$. However, if we define $\psi \diamond\!\!\rightarrow \phi$ as $\neg(\psi \rightsquigarrow \neg\phi)$, (CEM) forces us to admit that $\psi \rightsquigarrow \phi$ is equal to $\psi \diamond\!\!\rightarrow \phi$. One may defend (CEM) by altering the meaning of the "might be" conditional. In [Pollock 1976], Pollock gives a persuasive argument in favor of this definition of the "might be" conditional.

$(\psi \wedge \phi) \rightarrow (\psi \rightsquigarrow \phi)$. The new system is called **VC** .

8.3.2 Reasoning about Actions

One can define agents by their ability to take actions. Actions move agents from a (set of) possible world(s) to another (set of) possible world(s). We will use a formal language to describe the possible worlds and actions. It seems to be convenient for reasoning about actions to make a distinction between formulas which describe the possible worlds and formulas which describe the actions. However, we also have to specify the connection between a single action, represented by an action formula, and its partial preconditions and postconditions, represented by the general formulas.

Now, suppose we have a set of formulas S which denotes the actual world, and consider an action a with its partial precondition ρ and its partial postcondition ψ. Because of the problems of qualification and ramification in reasoning about actions [Ginsberg&Smith 1987], we may not be able to offer a complete description of the precondition and postcondition of each action. We have a set of formulas which we wish to preserve because those formulas represent laws or law-like statements, i.e., conditions that should not be violated, lest our logic becomes counterintuitive. We call this set of the formulas *protected sentences*, written S^*.

The key task in reasoning about action is to compute the result from applying action a to the initial world S. For an arbitrary action a, the postcondition ψ is generally not true in the actual world S [3]. If the counterfactual $\psi \rightsquigarrow \phi$ is true in the actual world S, this means that ϕ is true in the new world(s) S' resulting from action a. We can see that there is a close relationship between action logic and conditional logic, i.e., conditional logic can serve as a kind of action logic. [Jackson 1989] virtually identifies action logic and counterfactual logic. In [Ginsberg 1986], Ginsberg proposes several applications for action logic as conditional logic.

Conditional logic differs from action logic in its perspective. For example, the counterfactual "If the gravitational constant were to take on a slightly higher value in the immediate vicinity of the earth, then people would suffer bone fractures more frequently" is an admissible statement in counterfactual logic, but it would make little sense in action logic. One way to understand this is the following: because the premise is incompatible with physical law, no action could effect the consequent. Conversely, counterlegal action (action that would violated protected sentences) would take us into worlds which are inconsistent with the protected sentences.

8.3.3 Update

The update $\phi \circ \psi$ is closely related to the intensional conditional $\rightsquigarrow$ ("wiggle")[4] known from Stalnaker's and D. Lewis's work. One can express (although not term-define) the wiggle in terms of the update operator via the so-called *Ramsey-rule*

[3] One could even make the point that an action is only justified if its postcondition does not hold in the actual world.

[4] In fact, in Stalnaker, the "wiggle" is a "corner" ($>$); we prefer the wiggle because it frees the $>$ for other uses.

[Grahne 1991]:

$$\vdash (\chi \rightarrow (\phi \rightsquigarrow \psi)) \Leftrightarrow \vdash ((\chi \circ \phi) \rightarrow \psi).$$

The interpretation of $\phi \circ \psi$ is:

$$[\![\phi \circ \psi]\!]_M = \{w \in W : \exists w' \in W(w' \in [\![\phi]\!]_M \text{ and } w \in cw(w', [\![\psi]\!]_M))\},$$

which yields the set of worlds where ψ holds so that one could have gotten there from a closest ϕ world. Note that $\phi \circ \psi$ is a backward looking operator [Grahne 1991]. Therefore, the constraints of the closest world function for update operation are somewhat different from those for conditionals. The constraints for update are (CS1), (CS2), and (CSC) which is:

(CSC) $cw(w, [\![\phi]\!]_M) \cap [\![\psi]\!]_M \subseteq cw(w, [\![\phi \wedge \psi]\!]_M)$.

(CS1) ensures that the closest ϕ-worlds (relative to a given world) are indeed ϕ-worlds; (CS2) ensures that w is its own (and unique) closest ϕ-world if ϕ is true at w. (CSC) says that if ψ is true at the closest ϕ-world, then the closest ϕ-world is also a closest ϕ-and-ψ-world.

Chapter 9

Preference Logics

9.1 Preferences

Preferences play a crucial role in standard conceptualizations of rational action in economics and other social sciences [French 1988]. Formal AI took a different route by using "goals", rather than preferences, as the fundamental notion in action logics. Yet the goal notion might create problems, since goals are typically context-dependent, whereas goal-statements are typically context-independent (as already observed by H.A. Simon [Simon 1964]). Two of these problems are the counterintuitive necessitation rule for goals (every theorem must be a goal), and the equally counterintuitive closure of goals under logical implication (everything that is logically implied by a goal is a goal) [Cohen&Levesque 1987, Cohen&Levesque 1990, Rao&Georgeff 1991]. A preference-oriented action logic may avoid those problems.[1]

As mentioned above, an agent with "perfect" rationality would know all behavior alternatives and the exact outcome of each alternative; the agent would also have a complete preference ordering for those outcomes. An agent with bounded rationality, in contrast, may not know all alternatives, nor may she know the exact outcome of each; also, the agent may not have a complete preference ordering for those outcomes.

Possible world semantics provides a natural setting for modeling bounded rationality. We assume that propositions stand for the set of possible worlds where they are true, and that actions propel agents from (sets of) possible worlds to (sets of) possible worlds, so each primitive action constitutes an accessibility relation. The perfectly rational agent would have a complete description of her actual world, a complete knowledge of all accessibility relations, and a complete preference ordering over possible worlds. (For such an agent, decision making boils down to planning the route to the most preferred accessible world(s)). Boundedly rational agents, in contrast, may not know exactly where they are, what they can do, nor what they want. They may not have more than a partial description of their actual world (we call those partial descriptions *situations*, or *states* (of the world)), a partial knowl-

[1] For the details, see the section " goals" of the chapter "ALX3".

edge of the accessibility relations, and a partial preference ordering over conceivable situations. Be this as it may, the context of bounded rationality seems to require a justification of rational decisions in terms of preferences.

As discussed in [von Wright 1963] and elsewhere, there are many ways of stating preferences. We agree with von Wright who assumes that most, if not all, ways of stating preferences can be represented in a possible world framework, so that the preference statement becomes a statement about states of the world. For example, "I prefer taking a taxi to taking the bus" may be interpreted as "I prefer the state in which I am taking the taxi to the state in which I am taking the bus". Formally, we use the symbol $\boldsymbol{P}$ to denote the binary preference relation. $\phi \boldsymbol{P} \psi$ is read as "the state of affairs ϕ is preferable to the state of affairs ψ".

If preferences are stated relative to possible worlds, one can assume that an agent who states $\phi \boldsymbol{P} \psi$ should prefer a change to the state $\phi \wedge \neg\psi$ to a change to the state $\psi \wedge \neg\phi$. Following v. Wright [von Wright 1963], we call this assumption the *conjunction expansion principle*, and restrict attention to preference statements that obey it. We focus on a single agent and avoid temporal references in this chapter. This allows us to omit agent and time indexes for the preference operator. Also, in this chapter, we will not deal with preferences about preferences (e.g., preferences implied by statements such as "I'd rather prefer not to be a smoker")

Preferences of an agent may be more or less contingent upon her actual situation. Building a hierarchy of increasing contingency, we distinguish between *absolute preferences, preferences ceteris paribus, conditional preferences*, and *actual preferences.*

An *absolute preference* of ϕ over ψ should mean that every state of the world which contains ϕ and $\neg\psi$ is preferred to every state of the world which contains $\neg\phi$ and ψ. There can be at most one absolutely preferred proposition in a consistent preference order [von Wright 1963] (cf. the theorem 9.2.7). Examples of an absolute preference are hard to come by, but postmodern readers may think of Schwarzenegger's Terminator One as enacting an absolute preference. We use $\phi P_a \psi$ to denote the absolute preference of ϕ over ψ.

One may speak of a preference *ceteris paribus* of ϕ over ψ iff the agent favors a change to $\phi \wedge \neg\psi$ over a change to $\neg\phi \wedge \psi$, irrespective of what the state of the world is, but assuming that the world does not change in other features beside ϕ and ψ. Ceteris paribus preferences are more exacting than one may think. Assume, for example, that the agent prefers having the flu to having cancer ceteris paribus. Then she would have to stick to her preference even in worlds where there is a perfect cure for cancer but flu is terminal. We use $\phi P_{cp} \psi$ to denote the preference *ceteris paribus* of ϕ over ψ.

Conditional preferences are contingent upon a situation, i.e., a partial description of the world. One may say that an agent conditionally prefers ϕ to χ (with respect to ψ), if the agent prefers ceteris paribus $\phi \wedge \neg\chi \wedge \psi$ to $\neg\phi \wedge \chi \wedge \psi$, but may not prefer ceteris paribus $\phi \wedge \neg\chi \wedge \neg\psi$ to $\neg\phi \wedge \chi \wedge \neg\psi$. One may assume that most casual preference statements are contingent upon specific situations. The preference for flu over cancer, for example, will hold in most situations, except in those where there is a perfect cure for cancer and none for the flu. We use $\phi P_{[\psi]} \chi$ to denote the

conditional preference of ϕ over χ with respect to situation ψ.

An agent's *actual preferences* should be contingent upon the world where they are stated [Hansson 1989]. The statement "I would prefer a banana right now" is apparently of that type. The agent may prefer a banana here and now, but she may prefer an orange later (perhaps for the simple reason that she already had a banana). Actual preferences involve statements about what is better in one (usually the present) state of the world, not about what would be better in any other state of the world. In general, by saying "ϕ is better than ψ" one asserts that ϕ is better than ψ in the present world, but one does not commit oneself to any claim that ϕ would be better than ψ in any other possible world. This does not rule out the possibility that ϕ would also be better than ψ in other possible worlds, but this possibility is not part of the statement about the actual preference. We use $\phi P_{ac} \psi$ to denote the actual preference of ϕ over ψ.

It is sometimes suggested that one should distinguish between *intrinsic* and *extrinsic* preferences [Chisholm&Sosa 1966a, von Wright 1963, von Wright 1972]. A preference for ϕ over ψ is said to be extrinsic if a (non-circular) *reason* can be given as to *why* ϕ is preferred to ψ. Otherwise, the preference is said to be intrinsic. Unfortunately, the term "reason" is somewhat ambivalent. A "reason" may be identified with the condition under which a preferences holds; alternatively, the reason may be a preference itself. The first type of extrinsic preference is, in fact, covered by the conditional preference relation; the second type of extrinsic preference requires nested preference statements, and is therefore ignored in this chapter.

9.2 A Preference Logic Based on the Notion of Minimal Change

The formal semantics of this section is based on the notion of minimal change.

9.2.1 Syntax

The alphabet consists of a finite set of lower case Latin symbols $p, q, r, ...$ with or without subscripts or superscripts to denote primitive propositions. Lower case Greek letters $\rho, \phi, \psi, ...$, with or without subscript or superscript denote propositional formulae built up from primitive proposition symbols and the Boolean connectives. In addition, we have the symbols P_a, P_{cp}, $P_{[\psi]}$, and P_{ac} to denote two-place preference relations for, respectively, absolute preferences, ceteris paribus preferences, conditional preferences (conditional on ψ) and actual preferences. The symbol $\boldsymbol{P}$ serves as a metavariable to stand for one of the symbols P_a, P_{cp}, $P_{[\psi]}$, and P_{ac}. Finally, there is the two place connective '$\rightsquigarrow$' to denote the intensional conditional.

Let Φ_0 be the set of the primitive propositions. The set of propositional formulae Γ_0 is the minimal set of formulae closed under the following syntactic rules:

(S1a) $p \in \Phi_0 \Rightarrow p \in \Gamma_0$.
(S1b) $\phi \in \Gamma_0, \psi \in \Gamma_0 \Rightarrow (\phi \wedge \psi) \in \Gamma_0$.
(S1c) $\phi \in \Gamma_0 \Rightarrow \neg\phi \in \Gamma_0$.

The set of preference formulae Γ_p is the minimal set of formulae closed under the following syntactic rules:

(S2a) $\phi \in \Gamma_0, \psi \in \Gamma_0 \Rightarrow \phi P_{ac} \psi \in \Gamma_p$.
(S2b) $\phi \in \Gamma_0, \psi \in \Gamma_0, \rho \in \Gamma_0 \Rightarrow \phi P_{[\rho]} \psi \in \Gamma_p$.
(S2c) $\phi \in \Gamma_0, \psi \in \Gamma_0 \Rightarrow \phi P_{cp} \psi \in \Gamma_p$.
(S2d) $\phi \in \Gamma_0, \psi \in \Gamma_0 \Rightarrow \phi P_a \psi \in \Gamma_p$.

The Language $\mathbf{L_{Pr}}$ for the logic of preference is the minimal set of formulas closed under the following syntactic rules:

(S3a) $\phi \in \Gamma_0 \Rightarrow \phi \in \mathbf{L_{Pr}}$.
(S3b) $\phi \in \mathbf{L_{Pr}} \Rightarrow \neg\phi \in \mathbf{L_{Pr}}$.
(S3c) $\phi \in \mathbf{L_{Pr}}, \psi \in \mathbf{L_{Pr}} \Rightarrow (\phi \wedge \psi) \in \mathbf{L_{Pr}}$.
(S3d) $\phi \in \mathbf{L_{Pr}}, \psi \in \mathbf{L_{Pr}} \Rightarrow (\phi \rightsquigarrow \psi) \in \mathbf{L_{Pr}}$.
(S3e) $\phi \in \Gamma_p \Rightarrow \phi \in \mathbf{L_{Pr}}$.

Boolean connectives such as $\rightarrow$ and $\vee$ are defined in terms of $\neg$ and $\wedge$ as usual.

9.2.2 Formal Semantics (MCP-Semantics)

As outlined above, the notion of minimal change serves as a point of departure for the formal semantics that we call MCP-Semantics (*M*inimal *C*hange semantics for *P*reference relations). A ϕ-world closest to world w is the world at which ϕ is true and which is more similar to w than any other world at which ϕ is true. A closest world may be not unique. We use a *closest world function cw* which assigns to each formula ϕ and each world w a set of closest ϕ-worlds to w. $[\![\phi]\!]_M$ denotes the set of all worlds where ϕ is true in the model M.

We assume a comparison relation $\succ$ over the set of possible worlds W that reflects the intrinsic preferences of a particular agent. We assume that the comparison relation $\succ$ is irreflexive, transitive, and not empty. If $w_1 \succ w_2$, we say that the world w_2 is less preferred than w_1 (or conversely, that the world w_1 is more preferred than the world w_2). Sometimes we use $w_2 \prec w_1$ to denote $w_1 \succ w_2$. Furthermore, we extend the relation $\succ$ on the possible world set W into a relation $\succ$ on the powerset of W, i.e., for possible world sets W_1, W_2, we use $W_1 \succ W_2$ to denote that $(\forall w_1 \in W_1)(\forall w_2 \in W_2)(w_1 \succ w_2)$ and $W_1 \neq \emptyset$ and $W_2 \neq \emptyset$. It is easy to see that the comparison relation $\succ$ on the powerset of W has the following properties:

(i) Irreflexivity: $(X \not\succ X)$.
(ii) Normality: $(X \not\succ \emptyset)$, $(\emptyset \not\succ X)$.
(iii) Asymmetry: $X \succ Y \Rightarrow Y \not\succ X$.

(iv) Transitivity: $X \succ Y$ and $Y \succ Z \Rightarrow X \succ Z$.
(v) Quasi-Monotonicity: $X \succ Y$ and $X' \subset X$ and $Y' \subset Y$ and $X' \neq \emptyset$ and $Y' \neq \emptyset \Rightarrow X' \succ Y'$.

Formally, a MCP-semantic model for $\mathbf{L_{Pr}}$ is a tuple $M = \langle W, cw, \succ, V\rangle$, where W is a set of possible worlds, V is a valuation function as usual, $\succ \subseteq W \times W$ is a comparison relation, and cw is a closest world function which satisfies the constraints (CS1)-(CS5), and (CSN):

$$(CSN) \quad [\![\phi]\!]_M \neq \emptyset \Rightarrow cw(w, [\![\phi]\!]_M) \neq \emptyset.$$

(If a formula ϕ is true in some world, then for any world w, the set of ϕ-closest worlds to w is not empty.[2])

The truth conditions are as follows:
For the model M and a possible world w,

(T1) $M, w \models p$ where p is an atomic proposition, iff $w \in V(p)$.
(T2) $M, w \models \neg\phi$ iff $M, w \not\models \phi$.
(T3) $M, w \models \phi \wedge \psi$ iff $M, w \models \phi$ and $M, w \models \psi$.
(T4) $M, w \models \phi \rightsquigarrow \psi$ iff $cw(w, [\![\phi]\!]_M) \subseteq [\![\psi]\!]_M$.

For defining the truth conditions of the four different preference relations, we combine the conjunction expansion principle with the notion of minimal change in various ways. For defining actual preferences, we condition the conjunction expansion principle on the actual world w, since the agent's actual preferences are supposed to be contingent on the actual world:

(T5.1) $M, w \models \phi P_{ac} \psi$ iff $cw(w, [\![\phi \wedge \neg\psi]\!]_M) \succ cw(w, [\![\psi \wedge \neg\phi]\!]_M)$.

For a *conditional* preference of ϕ to ψ with respect to the condition ρ we require that the agent prefers ceteris paribus $\phi \wedge \neg\psi \wedge \rho$ to $\neg\phi \wedge \psi \wedge \rho$, but may not prefer ceteris paribus $\phi \wedge \neg\psi \wedge \neg\rho$ to $\neg\phi \wedge \psi \wedge \neg\rho$. So, the truth condition parallels (T5.2) with ρ added as the specific condition:

(T5.2) $M, w \models \phi P_{[\rho]} \psi$ iff $(\forall w' \in W) cw(w', [\![\phi \wedge \neg\psi \wedge \rho]\!]_M) \succ cw(w', [\![\psi \wedge \neg\phi \wedge \rho]\!]_M)$.

For the preference *ceteris paribus* of ϕ over ψ, we require that the agent favors a change to $\phi \wedge \neg\psi$ over a change to $\neg\phi \wedge \psi$, irrespective of what the state of the world is, but assuming that the world does not change in other features beside ϕ and ψ, hence:

(T5.3) $M, w \models \phi P_{cp} \psi$ iff $(\forall w' \in W) cw(w', [\![\phi \wedge \neg\psi]\!]_M) \succ cw(w', [\![\psi \wedge \neg\phi]\!]_M)$.

[2]We'll use this condition to prove some property involves the independency, which is discussed later in this chapter.

Finally, for an *absolute* preference, no reference to minimal change is required since an absolute preference of ϕ over ψ should mean that every state of the world which contains ϕ and $\neg\psi$ is preferred to every state of the world which contains $\neg\phi$ and ψ:

(T5.4) $M, w \models \phi P_a \psi$ iff $[\![\phi \wedge \neg\psi]\!]_M \succ [\![\psi \wedge \neg\phi]\!]_M$.

The MCP-semantics makes the logic of preferences *extensional*, so that preference relations are closed under the substitution of provable equivalents[3]. Formally, we have the following rules:

$$(PL) \quad \vdash \phi \leftrightarrow \chi \Rightarrow \vdash \phi \boldsymbol{P} \psi \leftrightarrow \chi \boldsymbol{P} \psi.$$

$$(PR) \quad \vdash \psi \leftrightarrow \chi \Rightarrow \vdash \phi \boldsymbol{P} \psi \leftrightarrow \phi \boldsymbol{P} \chi.$$

9.2.3 An Axiomatic Characterization of Preference Relations

Preferences have proven to be difficult to characterize syntactically. There is little consensus among logicians about the basic principles that guide a preference relation. Intuitively straightforward principles such as contraposition and conjunction expansion have been marred by counterexamples or worse. [Halldén 1966, Hansson 1968, Rescher 1967, Rescher 1967a, von Wright 1963, von Wright 1972]. Fortunately, much of the confusion goes away if appropriate distinctions between different preference relations are taken into consideration.

At the level of the formal semantics, we have distinguished four kinds of preference relations, ordered according to decreasing strength. An absolute preference of ϕ over ψ implies a ceteris paribus preference of ϕ over ψ, a ceteris paribus preference of ϕ over ψ implies a conditional preference ϕ over ψ, and so on. So, $M, w \models \phi P_a \psi \Rightarrow M, w \models \phi P_{cp} \psi \Rightarrow M, w \models \phi P_{[\top]} \psi \Rightarrow M, w \models \phi P_{ac} \psi$. This is borne out by the formal semantics, since the set of all ϕ-and $\neg\psi$ worlds is a superset of all ϕ-and $\neg\psi$ worlds closest to a particular arbitrary world, etc. This allows for a hierarchical characterization of preference relations, where the actual preference relation can be viewed as the primitive one. We consider the following axioms:[4]

[3] In [von Wright 1963], von Wright points out that expressions which are provably equivalent in the logic of propositions are not, without restriction, intersubstitutable in expressions of the logic of preference. The restriction is that the substitution must not introduce new variables. this restriction does not apply here

[4] For (UP) and (AP), we need the additional condition "α is independent of $\{\phi, \psi\}$". We shall give the details later in this section.

(N) $\neg(\bot \boldsymbol{P} \phi)$, $\neg(\phi \boldsymbol{P} \bot)$.
(AS) $\phi \boldsymbol{P} \psi \rightarrow \neg(\psi \boldsymbol{P} \phi)$.
(CEP) $\phi \boldsymbol{P} \psi \leftrightarrow (\phi \wedge \neg\psi)\boldsymbol{P}(\neg\phi \wedge \psi)$.
(COP) $\phi \boldsymbol{P} \psi \rightarrow (\chi \rightsquigarrow \phi \boldsymbol{P} \psi)$.
(UP) $\phi \boldsymbol{P} \psi \rightarrow ((\phi \wedge \alpha)\mathbf{P}(\psi \wedge \alpha) \wedge (\phi \wedge \neg\alpha)\boldsymbol{P}(\psi \wedge \neg\alpha))$.
(AP) $\phi \boldsymbol{P} \psi \rightarrow (\phi \wedge \neg\psi \wedge \alpha)\boldsymbol{P}(\psi \wedge \neg\phi \wedge \neg\alpha)$.

The first three axioms[5] – (N), (AS), (CEP) – together with the system **VC** and the inference rules (PL) and (PR) characterize actual preferences. Call this system **P1**. Adding (COP) to **P1** yields a characterization of conditional preferences (**P2**). **P2** and (UP) together characterize ceteris paribus preferences (**P3**). Adding finally (AP) to **P3** yields a charaterization of absolute preferences (**P4**).

We start with an axiomatic characterization of actual preferences and show then how this characterization extends to the other preference relations.

Actual Preferences: The System P1

We show the validity of asymmetry, contraposition, conjunction expansion, and irreflexivity on the class of MCP-models of actual preference (T1-4) and (T5.1).

9.2.1. Theorem. **P1** *is valid on the class of models of actual preference.*

PROOF. Straightforward from the properties of $\succ$ on the powerset of W. □

9.2.1. Proposition. (More properties of preference)
(CP) $\phi \boldsymbol{P} \psi \leftrightarrow (\neg\psi)\boldsymbol{P}(\neg\phi)$.
(IR) $\neg(\phi \boldsymbol{P} \phi)$.
(NT) $\neg(\top \boldsymbol{P} \phi)$, $\neg(\phi \boldsymbol{P} \top)$.

PROOF.
(CP) $\phi \boldsymbol{P} \psi \leftrightarrow (\neg\psi)\boldsymbol{P}(\neg\phi)$.

	$\vdash \phi \boldsymbol{P} \psi$	
$\Leftrightarrow$	$\vdash (\phi \wedge \neg\psi)\boldsymbol{P}(\psi \wedge \neg\phi)$	(CEP)
$\Leftrightarrow$	$\vdash (\neg\psi \wedge \neg(\neg\phi))\boldsymbol{P}(\neg\phi \wedge (\neg(\neg\psi))$	(Rules (PL) and (PR))
$\Leftrightarrow$	$\vdash \neg\psi \boldsymbol{P} \neg\phi$	(CEP)

(IR) $\neg(\phi \boldsymbol{P} \phi)$.

[5]Note that the transitivity axiom is missing from this list. We discuss the reason later.

$$\begin{array}{lll} & \vdash (\phi \boldsymbol{P} \phi) & \\ \Rightarrow & \vdash (\phi \wedge \neg\phi) \boldsymbol{P} (\phi \wedge \neg\phi) & \text{(CEP)} \\ \Rightarrow & \vdash \bot \boldsymbol{P} \bot & \text{(Definition of } \bot\text{)} \\ \Rightarrow & \vdash \bot & \text{(N)} \end{array}$$

So $\vdash (\phi \boldsymbol{P} \phi) \rightarrow \bot$. Therefore, $\vdash \neg(\phi \boldsymbol{P} \phi)$.

(NT) $\neg(\top \boldsymbol{P} \phi)$, $\neg(\phi \boldsymbol{P} \top)$.

$$\begin{array}{lll} & \vdash (\top \boldsymbol{P} \phi) & \\ \Rightarrow & \vdash (\neg\phi) \boldsymbol{P} (\neg\top) & \text{(CP)} \\ \Rightarrow & \vdash (\neg\phi) \boldsymbol{P} \bot & \text{(Definition of } \top\text{)} \\ \Rightarrow & \vdash \bot & \text{(N)} \end{array}$$

So, $\vdash \neg(\top \boldsymbol{P} \phi)$. The proof for the second half of (NT) is similar. □

Conditional Preferences: The System P2

We show the validity of the *conditionality principle* (COP) on the class of models of conditional preference (T1-4) and (T5.2):

$$(COP) \quad \phi \boldsymbol{P} \psi \rightarrow (\chi \rightsquigarrow \phi \boldsymbol{P} \psi).$$

The conditionality principle provides the watershed between actual and conditional preferences, because it states that the latter are independent of minimal changes. It characterizes the relation between conditional preferences and an (arbitrary) action in terms of the action's weakest postcondition, χ. An arbitrary action cannot cause a change in the conditional preferences of an agent, because it does not affect that condition (as opposed to actual preferences that may be different at different worlds).

9.2.2. Proposition. *(COP) is valid on the class of models of conditional preference.*

PROOF. (T5.2) implies that $\phi P_{[\rho]} \psi$ is true at all worlds if it is true at all, i.e., for any MCP model $M = \langle W, cw, \succ, V \rangle$, and any world $w \in W$, if $M, w \models \phi P_{[\rho]} \psi$, then $M, w' \models \phi P_{[\rho]} \psi$ for any $w' \in W$, so that $W = [\![\phi P_{[\rho]} \psi]\!]_M$. Trivially, for any $\chi, cw(w, [\![\chi]\!]_M) \subseteq W$, hence, due to the truth conditions for the conditional $M, w \models \chi \rightsquigarrow (\phi P_{[\rho]} \psi)$ holds for arbitrary w. So: $\phi P_{[\rho]} \psi \rightarrow (\chi \rightsquigarrow \phi P_{[\rho]} \psi)$, and therefore $\phi \boldsymbol{P} \psi \rightarrow (\chi \rightsquigarrow \phi \boldsymbol{P} \psi)$ for the class of models of conditional preference. □

We call the logic system which consists of **P1** and (COP) the *logic of conditional preference*, abbreviated **P2**.

Preferences Ceteris Paribus: The System P3

We show the validity of the *unconditionality principle* on the class of models of unconditional preference (T1-4) and (T5.3). (UP) provides the watershed between preferences *ceteris paribus* and *conditional* preferences, because it requires that for each condition α, a preference of ϕ over ψ holds both for the condition α and its negation, $\neg\alpha$. However, we must require independence for the formula α to avoid preference statements with the *falsum* as one of the arguments, since such statements would lead to a violation of the irreflexivity of $\boldsymbol{P}$; also, such statements would be hard to interpret intuitively.

We define α is independent of $\{\phi, \psi\}$ as follows:

9.2.3. Definition. (Independence) $IND(\alpha, \{\phi, \psi\}) \overset{\text{def}}{\Longleftrightarrow}$
$(\phi \wedge \neg\psi \diamond\!\!\rightarrow \alpha) \wedge (\psi \wedge \neg\phi \diamond\!\!\rightarrow \alpha) \wedge (\phi \wedge \neg\psi \diamond\!\!\rightarrow \neg\alpha) \wedge (\psi \wedge \neg\phi \diamond\!\!\rightarrow \neg\alpha)$,
where $\diamond\!\!\rightarrow$ *is the "might-be" conditional, which is defined as:*
$\phi \diamond\!\!\rightarrow \psi \overset{\text{def}}{\Longleftrightarrow} \neg(\phi \rightsquigarrow \neg\psi)$.

9.2.4. Lemma. (Independency lemma) *For any MCP model M, and any world w,* $M, w \models IND(\alpha, \{\phi, \psi\}) \Rightarrow [\![\alpha \wedge \phi \wedge \neg\psi]\!]_M \neq \emptyset$.

PROOF.

	$M, w \models IND(\alpha, \{\phi, \psi\})$	
$\Rightarrow$	$M, w \models \phi \wedge \neg\psi \diamond\!\!\rightarrow \alpha$	(Definition of IND)
$\Rightarrow$	$M, w \models \neg(\phi \wedge \neg\psi \rightsquigarrow \neg\alpha)$	(Definition of $\diamond\!\!\rightarrow$)
$\Rightarrow$	$cw(w, [\![\phi \wedge \neg\psi]\!]_M) \cap [\![\alpha]\!]_M \neq \emptyset$	(Truth condition)
$\Rightarrow$	$[\![\phi \wedge \neg\psi]\!]_M \cap [\![\alpha]\!]_M \neq \emptyset$	(CS1)
$\Rightarrow$	$[\![\phi \wedge \neg\psi \wedge \alpha]\!]_M \neq \emptyset$	(Truth condition)

□

So the axiom (UP) becomes:

(UP) $IND(\alpha, \{\phi, \psi\}) \rightarrow (\phi \boldsymbol{P} \psi \rightarrow (\phi \wedge \alpha) \boldsymbol{P} (\psi \wedge \alpha) \wedge (\phi \wedge \neg\alpha) \boldsymbol{P} (\psi \wedge \neg\alpha))$.

9.2.5. Proposition. *The unconditionality principle (UP) is valid on the class of models of ceteris paribus preference for independent conditions α.*

PROOF. Suppose that α is independent of $\{\phi, \psi\}$. By the lemma, we have $[\![\phi \wedge \neg\psi \wedge \alpha]\!]_M \neq \emptyset$. Moreover, by (CSN), we have $cw(w, [\![\phi \wedge \neg\psi \wedge \alpha]\!]_M) \neq \emptyset$ for any world w. Similarly, we have $cw(w, [\![\psi \wedge \neg\phi \wedge \alpha]\!]_M) \neq \emptyset$.

Furthermore, from $M, w \models IND(\alpha, \{\phi, \psi\})$, we have $cw(w, [\![\phi \wedge \neg\psi]\!]_M) \cap [\![\alpha]\!]_M \neq \emptyset$. Therefore, by (CS5), we have $cw(w, [\![\phi \wedge \neg\psi \wedge \alpha]\!]_M) \subseteq cw(w, [\![\phi \wedge \neg\psi]\!]_M)$. Similarly, we have $cw(w, [\![\psi \wedge \neg\phi \wedge \alpha]\!]_M) \subseteq cw(w, [\![\psi \wedge \neg\phi]\!]_M)$.

(T5.3) assures that $M, w' \models \phi \boldsymbol{P} \psi \Leftrightarrow cw(w, [\![\phi \wedge \neg\psi]\!]_M) \succ cw(w, [\![\psi \wedge \neg\phi]\!]_M)$ for all w, hence, $M, w' \models \phi \boldsymbol{P} \psi \Rightarrow cw(w, [\![\phi \wedge \neg\psi \wedge \alpha]\!]_M) \succ cw(w, [\![\psi \wedge \neg\phi \wedge \alpha]\!])$

for all w, by the quasi-monotonicity of $\succ$. By the same token, $M, w' \models \phi \boldsymbol{P} \psi \Rightarrow cw(w, [\![\phi \wedge \neg\psi \wedge \neg\alpha]\!]_M) \succ cw(w, [\![\psi \wedge \neg\phi \wedge \neg\alpha]\!]_M)$ for all w, so that the truth conditions of $(\phi \wedge \alpha)\boldsymbol{P}(\psi \wedge \alpha)$ and $(\phi \wedge \neg\alpha)\boldsymbol{P}(\psi \wedge \neg\alpha))$ are fulfilled. □

We call the logic system which consists of **P2** and (UP) the *logic of preference ceteris paribus*, written **P3**.

Absolute Preferences: The System P4

We show the validity of the *absolute preference* principle (AP) on the class of models of absolute preference (T1-4) and (T5.4):

Again, we must require that α is independent of $\{\phi, \psi\}$.

$$(AP) \quad IND(\alpha, \{\phi, \psi\}) \rightarrow (\phi \boldsymbol{P} \psi \rightarrow (\phi \wedge \neg\psi \wedge \alpha)\boldsymbol{P}(\psi \wedge \neg\phi \wedge \neg\alpha)).$$

(AP) provides the watershed between preferences ceteris paribus and absolute preferences by expressing the independence of a given preference for any arbitrary condition.

9.2.6. Proposition. *(AP) is valid on the class of models of absolute preference.*

PROOF. $M, w \models IND(\alpha, \{\phi, \psi\}) \wedge \phi P_a \psi$
$\Rightarrow [\![\phi \wedge \neg\psi \wedge \alpha]\!]_M \neq \emptyset$ and $[\![\psi \wedge \neg\phi \wedge \neg\alpha]\!]_M \neq \emptyset$ and $[\![\phi \wedge \neg\psi]\!]_M \succ [\![\psi \wedge \neg\phi]\!]_M$
$\Rightarrow [\![\phi \wedge \neg\psi \wedge \alpha]\!]_M \succ [\![\psi \wedge \neg\phi \wedge \neg\alpha]\!]_M$
$\Rightarrow [\![\phi \wedge \neg\psi \wedge \alpha \wedge \neg(\psi \wedge \neg\phi \wedge \neg\alpha)]\!]_M \succ [\![\psi \wedge \neg\phi \wedge \neg\alpha \wedge \neg(\phi \wedge \neg\psi \wedge \alpha)]\!]_M$
$\Rightarrow M, w \models (\phi \wedge \neg\psi \wedge \alpha) P_a (\psi \wedge \neg\phi \wedge \neg\alpha)$. □

We call the logic system which consists of **P3** and (AP) the *logic of absolute preference*, written **P4**.

It is non-trivial to see that there can be only one absolute preference in a consistent preference order. In [von Wright 1963], von Wright claims the following statement without proof.

9.2.7. Theorem. (von Wright) *There exists only one absolute preference in a consistence preference order, i.e.,*
$(\phi P_a \psi \wedge \rho P_a \chi) \Rightarrow$ **False**, *if* $\chi \wedge \neg\rho$ *and* $\rho \wedge \neg\chi$ *are independent of* $\{\phi, \psi\}$.

PROOF. $\chi \wedge \neg\rho$ and $\neg\chi \wedge \rho$ are independent of $\{\phi, \psi\}$
$\Rightarrow [\![\psi \wedge \neg\phi \wedge \rho \wedge \neg\chi]\!]_M \neq \emptyset$ and $[\![\rho \wedge \neg\chi \wedge \psi \wedge \neg\phi]\!]_M \neq \emptyset$ and
$[\![\phi \wedge \neg\psi \wedge \chi \wedge \neg\rho]\!]_M \neq \emptyset$ (by the independency lemma)

Therefore, $M, w \models (\phi P_a \psi) \wedge (\rho P_a \chi)$
$\Rightarrow [\![\phi \wedge \neg\psi]\!]_M \succ [\![\psi \wedge \neg\phi]\!]_M$ and $[\![\rho \wedge \neg\chi]\!]_M \succ [\![\chi \wedge \neg\rho]\!]_M$ (by the truth condition)
$\Rightarrow [\![\phi \wedge \neg\psi \wedge \chi \wedge \neg\rho]\!]_M \succ [\![\psi \wedge \neg\phi \wedge \rho \wedge \neg\chi]\!]_M$ and $[\![\rho \wedge \neg\chi \wedge \psi \wedge \neg\phi]\!]_M \succ [\![\chi \wedge \neg\rho \wedge \phi \wedge \neg\psi]\!]_M$
(Because $[\![\phi \wedge \neg\psi]\!]_M \supseteq [\![\phi \wedge \neg\psi \wedge \chi \wedge \neg\rho]\!]_M$, etc.)
$\Rightarrow [\![\phi \wedge \neg\psi \wedge \chi \wedge \neg\rho]\!]_M \succ [\![\phi \wedge \neg\psi \wedge \chi \wedge \neg\rho]\!]_M$ (Because of the transitivity of $\succ$)
$\Rightarrow$**False** (Because the irreflexivity of $\succ$) □

9.3 From Preference Statements to Preferences on Possible Worlds

The characterization of the four preference relations enables us to reverse the emphasis now, discussing briefly how preference statements may be transformed into preferences over possible worlds. One will have to do that in one way or another when one wants to use preferences in the context of an action logic. Recall the informal setting from the introduction. We assume a boundedly rational agent who may have a partial description of her actual world, a partial knowledge of accessibility relations, and a partial preference ordering over conceivable states of affairs. States of affairs are equivalent to propositions. So, agents may know, for example, that they prefer ϕ to ψ ceteris paribus, or that they do so conditionally for situation ρ, or that they do so only in the actual world. Yet they may not have a complete preference order over possible worlds – in fact, they are quite unlikely to have such an order, unless they are omniscient or, alternatively, possible worlds are conceived in a very restricted language (as in the example below). When making a decision about a course of action, the rational agent will have to determine the set of most preferred, but accessible worlds. So, given a preference ordering on state of affairs, we may want to know the corresponding ordering on (sets of) possible worlds. We can use a *world lattice* to represent the possible worlds and their closest world relations. Let Φ_0 be the primitive proposition set. Then a world lattice L is a lattice $\langle W, \leq \rangle$, where $W = \mathcal{P}(\Phi_0)$, is called possible world set, and $\langle w, w' \rangle \in \leq$ iff $w \subseteq w'$. Moreover, we view each possible world as a mapping from the primitive proposition set to the set $\{0, 1\}$. Namely, $w(p_i) = 1$ iff $p_i \in w$, $w(p_i) = 0$ iff $p_i \notin w$, for $p_i \in \Phi_0$. We define the distance between two worlds as follows:

$$d(w, w') \stackrel{\text{def}}{=} |\{p \in \Phi_0 : w(p) \neq w'(p)\}|.$$

For a world lattice $L = \langle W, \leq \rangle$, we define a minimal change model M, which is called a *minimal change model induced from the world lattice* L, $M = \langle W, cw, V \rangle$ as follows:

$w \in V(p_i)$ iff $p_i \in w$. Namely, $M, w \models p_i$ iff $p_i \in w$, for any $p_i \in \Phi_0$.
$w' \in cw(w, [\![\varphi]\!]_M)$ iff (i) $M, w' \models \varphi$ and (ii) there exists no other $w" \in W$ such that $M, w" \models \varphi$ and $d(w, w") < d(w, w')$.

For example, if we assume that the set of atomic propositions consists of $\{p, q, r\}$, the corresponding world lattice is shown in Figure 9.1. If we assume for simplicity that there are no material axioms (that would make certain possible worlds infeasible) then the nearest worlds in the world lattice are the closest worlds. For example, suppose the present world w is $\{p, q, r\}$, then the closest worlds $cw(w, [\![p \wedge \neg q]\!]_M)$ must be the world $\{p, r\}$, because the world $\{p, r\}$ is nearer to w than the world $\{p\}$. In other words, the world $\{p, r\}$ embodies fewer atomic changes from the present

world w than the world $\{p\}$. However, if there are material axioms, say, $p \rightarrow \neg r$ is a material axiom, then the closest (feasible) world becomes $\{p\}$ instead of $\{p, r\}$.

9.3.1. THEOREM. (Induced minimal change model theorem) *For any world lattice L, the induced minimal change model M satisfies (CS1)-(CS5) and (CSN).*

PROOF. (CS1) if $w' \in cw(w, [\![\psi]\!]_M)$ then $w' \in [\![\psi]\!]_M$;
straightforward from the definition.

(CS2) if $w \in [\![\psi]\!]_M$ then $cw(w, [\![\psi]\!]_M) = \{w\}$;
Suppose $w \in [\![\psi]\!]_M$
(i) $\{p \in \Phi_0 : w(p) \neq w(p)\} = \emptyset$
$\Rightarrow d(w, w) = 0$
$\Rightarrow w \in cw(w, [\![\psi]\!]_M)$

(ii) suppose there exists another world x such that $x \in cw(w, [\![\phi]\!]_M)$, then $d(w, x) = 0$, that is, $x = w$.

Therefore, $cw(w, [\![\phi]\!]_M) = \{w\}$.

(CS3) if $cw(w, [\![\psi]\!]_M)$ is empty then $cw(w, [\![\phi]\!]_M) \cap [\![\psi]\!]_M$ is also empty;
$cw(w, [\![\psi]\!]_M) = \emptyset \Rightarrow [\![\psi]\!]_M = \emptyset \Rightarrow cw(w, [\![\phi]\!]_M) \cap [\![\psi]\!]_M = \emptyset$.

(CS4) if $cw(w, [\![\psi]\!]_M) \subseteq [\![\phi]\!]_M$ and $cw(w, [\![\phi]\!]_M) \subseteq [\![\psi]\!]_M$, then $cw(w, [\![\psi]\!]_M) = cw(w, [\![\phi]\!]_M)$;
For any $y \in cw(w, [\![\psi]\!]_M)$ and any $x \in cw(w, [\![\phi]\!]_M)$. Let $m = d(w, y)$, and $n = d(w, x)$, we argue that $m = n$.
$m < n \Rightarrow \exists z \in cw(w, [\![\psi]\!]_M)(d(w, z) < d(w, x))$
$\Rightarrow \exists z \in [\![\phi]\!]_M(d(w, z) < d(w, x))$, which contradicts with $x \in cw(w, [\![\phi]\!]_M)$. Similarly, we can show that $m > n$ also causes falsum. Therefore, we conclude that $m = n$.

$y \in cw(w, [\![\psi]\!]_M) \Rightarrow y \in [\![\phi]\!]_M$ and $d(w, y) = n$
$\Rightarrow y \in cw(w, [\![\phi]\!]_M)$. Therefore, $cw(w, [\![\psi]\!]_M) \subseteq cw(w, [\![\phi]\!]_M)$
Similarly, we can show that $cw(w, [\![\phi]\!]_M) \subseteq cw(w, [\![\psi]\!]_M)$. Therefore, $cw(w, [\![\psi]\!]_M) = cw(w, [\![\phi]\!]_M)$.

(CS5) if $cw(w, [\![\psi]\!]_M) \cap [\![\phi]\!]_M \neq \emptyset$, then $cw(w, [\![\psi \wedge \phi]\!]_M) \subseteq cw(w, [\![\psi]\!]_M)$.
Suppose that $cw(w, [\![\psi]\!]_M) \cap [\![\phi]\!]_M \neq \emptyset$, then for any $x \in W$,

$x \in cw(w, [\![\phi \wedge \psi]\!]_M)$ and $x \notin cw(w, [\![\phi]\!]_M)$
$\Rightarrow x \in [\![\phi]\!]_M$ and $x \in [\![\psi]\!]_M$ and $\forall(y \in [\![\phi \wedge \psi]\!]_M)(d(w, x) \leq d(w, y))$ and $\forall(z \in cw(w, [\![\phi]\!]_M))(z \in [\![\psi]\!]^v_M \Rightarrow (d(w, z) \leq d(w, x)))$
$\Rightarrow \forall(y \in [\![\phi \wedge \psi]\!]_M)(d(w, x) \leq d(w, y))$ and $\exists(z \in cw(w, [\![\phi]\!]_M) \cap [\![\psi]\!]_M)(d(w, z) \leq d(w, x))$
$\Rightarrow \forall(y \in [\![\phi \wedge \psi]\!]_M)(d(w, x) \leq d(w, y))$ and $\exists(z \in [\![\phi \wedge \psi]\!]_M)(d(w, z) \leq d(w, x))$
$\Rightarrow$**False**

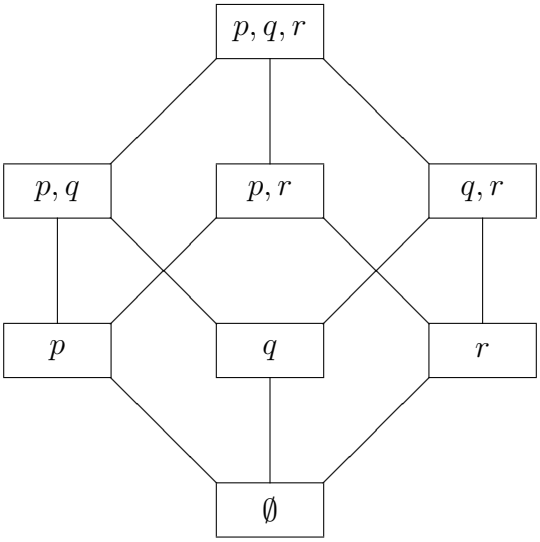


Figure 9.1: World Lattices

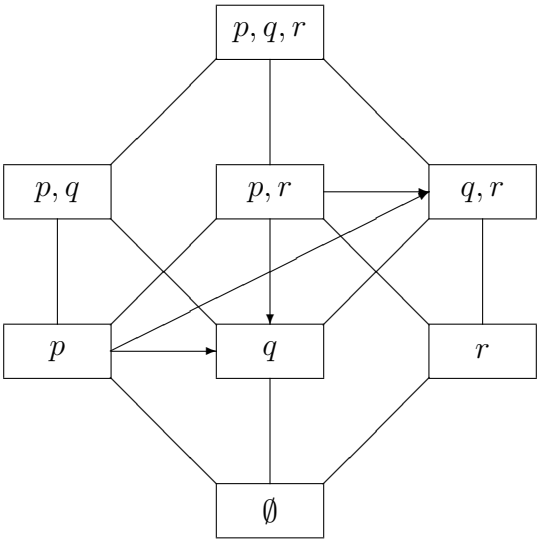
(CSN) if $[\![\psi]\!]_M \neq \emptyset$, then $cw(w, [\![\psi]\!]_M) \neq \emptyset$.
Suppose that $[\![\psi]\!]_M \neq \emptyset$. Consider the set D of distances between w and $w' \in [\![\psi]\!]_M$, namely, $D = \{d(w, w') : w' \in [\![\psi]\!]_M\}$. Let the set $MD = \{w" \in [\![\psi]\!]_M : d(w, w") = MIN(D)\}$, In other words, the elements in MD are those worlds $w" \in [\![\psi]\!]_M$ such that $w"$ is closest to w. It is easy to see that the set MD is not empty and $MD = cw(w, [\![\psi]\!]_M)$. So $cw(w, [\![\psi]\!]_M) \neq \emptyset$. □

For a preference statement pP_aq, we use $CR(pP_aq)$ to denote the partial comparison relation on the possible worlds that correspond to it. We can construct a general mapping $CR : \Gamma_p \to \mathcal{P}(W \times W)$ from preferences on state of affairs to a comparison relation on possible worlds such that for any MCP model $M = \langle W, cw, CR(\phi), V\rangle$, $M, w \models \phi$.
If we assume that ϕ and ψ are logically independent, the problem is simple. We just have the following:

$CR(\phi P_a \psi) = \{\langle w_1, w_2\rangle : w_1 \in [\![\phi \wedge \neg\psi]\!]_M$ and $w_2 \in [\![\neg\phi \wedge \psi]\!]_M\}$.
$CR(\phi P_{cp} \psi) = \{\langle w_1, w_2\rangle : \exists w(w_1 \in cw(w, [\![\phi \wedge \neg\psi]\!]_M)$ and $w_2 \in cw(w, [\![\neg\phi \wedge \psi]\!]_M))\}$.
$CR(\phi P_{[\rho]} \psi) = \{\langle w_1, w_2\rangle : \exists w(w_1 \in cw(w, [\![\phi \wedge \neg\psi \wedge \rho]\!]_M)$ and $w_2 \in cw(w, [\![\neg\phi \wedge \psi \wedge \rho]\!]_M))\}$.

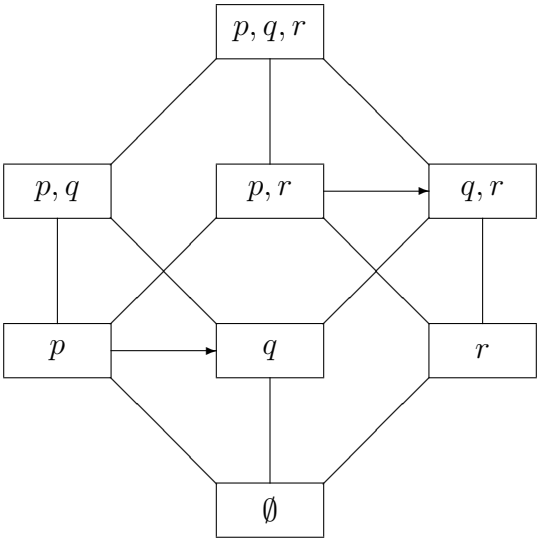
The relations concerning the absolute preference pP_aq, preference *ceteris paribus* $pP_{cp}q$, and conditional preference $pP_{[r]}q$ are shown in the figures.

Given a set of preference statements, material axioms, and accessibility relations, the agent may now uniquely determine a set of most preferred, but accessible possible



Absolute Preference pP_aq

Figure 9.2: Absolute Preference



Preference ceteris paribus $pP_{cp}q$

Figure 9.3: Preference Ceteris Paribus

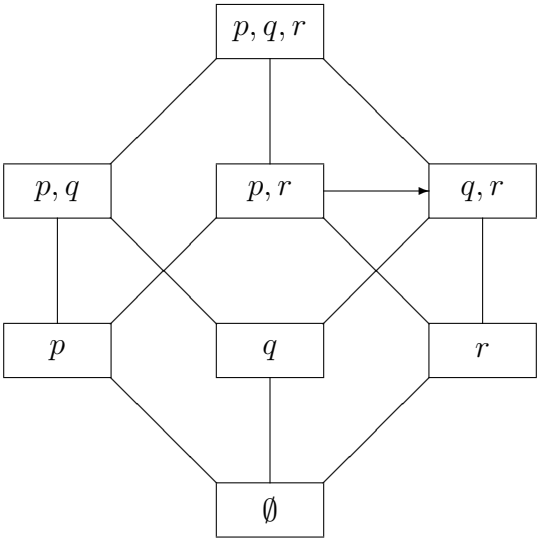


Figure 9.4: conditional Preference

worlds.

9.4 Discussion

We discuss briefly how our logic compares to other preference logics, reiterate some well-known counterexamples, and discuss the import of preference logic for action logic.

9.4.1 Other Approaches to Preference Logic

The major early contributions by [Halldén 1957] and [von Wright 1963] had no formal semantics. As in other areas of modal logic, this led to an inflation of syntactic approaches, so that by the mid-'60s a large number of axiomatic characterizations were competing for recognition. In his overview of preference logics, Rescher had to conclude in 1967 that practically all preference principles were contested.[6] Confusion, and then stagnation was the result; it lasted until the late '80s. The first formal semantics (by Rescher) was of little help [Rescher 1967a], possibly because it was based on the unrealistic assumption of numerical *measures of goodness* associated with possible worlds, and hence on a complete preference ordering.

There has been something of a renaissance of preference logic the last couple of years. Two new semantic approaches to preference logic were proposed by [Hansson 1989] and by [Brown et al. 1991]. Hansson differs from our approach in

[6]Rescher makes an exceptions for irreflexivity and transitivity, but there are also reflexive preference logics, e.g., [Hansson 1989] and transitivity was later attacked by [Mullen 1979].

using an essentially algebraic representation function to capture the comparison relation. His approach is also more restricted by imposing uniqueness for minimally different possible worlds. Hansson has no notion of conditional preference (although his approach could accommodate conditional preferences). Most importantly, he has no clear axiomatic characterization of different kinds of preference relations. We refer the reader to [Hansson 1989] for details. Still, both Hansson and our approach are squarely in the v. Wright tradition of two-place modalities for preference relations.

Brown, Mantha, and Wakayama [Brown et al. 1991] take a different tack by building a formal semantic of preferences on two one-place preference operators P_f and P_b. Roughly, $P_f F$ is true at a world w if all successor worlds where F is true are weakly preferred to w. Conversely, $P_b F$ is true at a world w if all successor worlds where F is true are not weakly preferred to w. This approach has certain important advantages. In particular, is has syntactic means of imposing restrictions on preference orderings among possible world, so that principles such as contraposition, conjunction expansion, or transitivity need not be built into the formal semantics. This makes their logic virtually counterexample-proof. Also, their approach easily accommodates nested preferences statements. We refer the reader to [Brown et al. 1991] for details. For two reasons, however, we did not follow Brown et al. in their approach. First, their semantics in not based on minimal change. As a consequence, it is not possible to express actual preferences in their logic. Also, it will be difficult, if not impossible, to use their logic in the context of an action logic that relies on minimal change as a means to solve the frame problem. Second, we see no way to express the conjunction expansion principle (the principle that embodies the essence of a possible world approach in our view). Even simple preference statements take on syntactically contrived shapes; $p\mathbf{P}q$ for example becomes $q \rightarrow P_f p$.

9.4.2 Counterexamples

As opposed to Brown et al., the semantics of this paper comes with some built-in properties for the preference relation, notably contraposition, conjunction expansion. As suggested before, those properties gave rise to counterexamples in the past. We discuss some of the best-known counterexamples in this section.

Contraposition. [Chisholm&Sosa 1966a, Chisholm&Sosa 1966b] propose the following counterexample: assume a hedonistic theory of intrinsic preferability and intrinsic value. Assume that a state of affairs is intrinsically good iff it entails the existence of more pleasure than displeasure (and conversely for bad states). Now, let p be "there being happy Americans" and q be "there being stones". The state "there being happy Americans" is intrinsically preferable to "there being stones", yet the contraposition does not seem to make sense – why should it be intrinsically preferable to have no stones rather than no happy Americans? We think that the counterexample falls apart once it is properly stated within the restrictions of a two-valued propositional language. A correct rendering of the statement would have to make explicit that the state of "there being stones" is devoid of happyness. Let a stand for Americans, h for happyness, and s for the presence of stones. We obtain $(a \wedge h)\mathbf{P}(s \wedge \neg h)$; its contraposition is $(\neg s \vee h)\mathbf{P}(\neg a \vee \neg h)$, expressing that we prefer

the absence of stones or happyness to the absence of Americans or unhappyness. This makes sense, since the propositions "there being stones" and "there being Americans" are now of equal intrinsic value (none), so that we prefer happyness to unhappyness. The counterexample rests on the implication of a third truth value while using a two-valued logic that cannot accommodate that truth value.

Hansson's counterexample against contraposition [Hansson 1968] assumes a person, A, who has bought some tickets in a lottery with two prizes of unequal worth. Let p stand for "A wins the first price" and q for "A wins some price". We may assume $p\mathbf{P}q$ for A but may not want to assume $\neg q\mathbf{P}\neg p$. We think that the counterexample works because the propositional language is again stretched beyond its limits – suggesting the dependence of p on q (winning the first price implies winning some price) but not making this dependence explicit. If one were to make this dependence explicit, one would have to state a preference for p and q over q and not p – and the contraposition makes sense again: preferring winning no price or winning the first price to winning no price or winning some (but possibly not the first) price.

Conjunction Expansion. Chisholm and Sosa also attack the conjunction expansion principle. Assume that it is better that Smith *and* his wife are happy ($p\wedge q$) than that Smith alone is happy. Conjunction expansion yields $(p\wedge q\wedge\neg p)\mathbf{P}\neg(p\wedge q)\wedge p$ and hence the preference for a contradictory state of affairs. As in earlier cases, the example stretches propositional logic beyond its limits. A correct rendering of the preference statement would have to make explicit that one prefers the state where both Smith and his wife are happy to the state where Smith is happy and his wife is not happy, i.e., $p \wedge q\mathbf{P}p \wedge \neg q$. But the conjunction expansion of $p \wedge q\mathbf{P}p \wedge \neg q$ is equivalent to $p \wedge q\mathbf{P}p \wedge \neg q$.

Another famous counterexample is by [Danielsson 1968]: one might assume that it is better that there is water in the swimming pool (w) and I do not jump into the pool ($\neg j$) than that there is no water in the swimming pool and I do jump into the pool. Because of the conjunction expansion, $(\neg j \wedge w\mathbf{P}j \wedge \neg w) \leftrightarrow w\mathbf{P}j$, one might want to conclude that it is better that there is water in the pool than that I jump into the pool. As in earlier cases, an implicature is used, but not made explicit, namely that jumping into an empty pool hurts. If this implicature is made explicit, (CEP) no longer produces counterintuitive results.

We may conclude that there is basically one source for unjustified counterexamples: the language of two-valued propositional language is stretched beyond its limits. Once this is corrected, the counterexamples wither away.[7]

9.5 Transitivity of Preferences

We have already noted that the transitivity, i.e., the axiom:

$$(TR) \quad \phi\boldsymbol{P}\psi \wedge \psi\boldsymbol{P}\rho \rightarrow \phi\boldsymbol{P}\rho,$$

[7]There are also attacks against other principles axiomatized in this thesis, i.e., transitivity and the unconditionality principle, but the attacks against transitivity relies on a notion of indifference that differs from ours [Mullen 1979], and the attack against the unconditionality principle does not take into account the requirement of independence [Hansson 1968]

is missing in the axiom system. Actually, the standard MCP semantics is not powerful enough to deal with the transitivity of preferences. Although the comparison relation on the possible world set W is transitive, the condition is still not enough to prove the transitivity, since the intersections of some closest world sets in questions may be empty. Here is a counterexample for the transitivity of actual preference.

9.5.1. CLAIM. *There exits a MCP model $M = \langle W, cw, \succ, V\rangle$ and a world $w \in W$ such that preferences are not transitive, even though the comparison relation $\succ$ is transitive. Especially, we claim that $M, w \models (p\boldsymbol{P}q) \wedge (q\boldsymbol{P}r) \wedge \neg(p\boldsymbol{P}r)$.*

PROOF. Suppose that the primitive proposition set is $\{p, q, r\}$. We define the model $M = \langle W, cw, \succ, V\rangle$ as follows:

$W = \{w_{pqr}, w_{pq}, w_{pr}, w_{qr}, w_p, w_q, w_r, w_\emptyset\}$.

We define cw (to the extent that we need for the example.)

$cw(w_p, [\![p \wedge \neg q]\!]_M) = \{w_p\}$.
$cw(w_p, [\![\neg p \wedge q]\!]_M) = \{w_q\}$.
$cw(w_p, [\![q \wedge \neg r]\!]_M) = \{w_{pq}\}$.
$cw(w_p, [\![\neg q \wedge r]\!]_M) = \{w_{pr}\}$.
$cw(w_p, [\![p \wedge \neg r]\!]_M) = \{w_p\}$.
$cw(w_p, [\![\neg p \wedge r]\!]_M) = \{w_r\}$.
.....
$\succ = \{< w_p, w_q >, < w_{pq}, w_{pr} >\}$.

$V(p) = \{w_{pqr}, w_{pq}, w_{pr}, w_p\}$.
$V(q) = \{w_{pqr}, w_{pq}, w_{qr}, w_q\}$.
$V(r) = \{w_{pqr}, w_{pr}, w_{qr}, w_r\}$.

It is easy to see that the above model M is a MCP model and $M, w_p \models (p\boldsymbol{P}q) \wedge (q\boldsymbol{P}r) \wedge \neg(p\boldsymbol{P}r)$. Therefore, the transitivity does not hold in the model M. Moreover, the transitivity of actual preference does not holds for MCP models. Note that in the model M the transitivity of the comparison relation $\succ$ is trivially true. □

In order to capture the transitivity axiom for preference, we have to change the MCP semantics. A convenient way to do this is to introduce the following additional condition in the semantic models:

(TRAN) $cw(w, X \cap \overline{Y}) \succ cw(w, Y \cap \overline{X})$ and $cw(w, Y \cap \overline{Z}) \succ cw(w, Z \cap \overline{Y}) \Rightarrow cw(w, X \cap \overline{Z}) \succ cw(w, Z \cap \overline{X})$.
where $\overline{Y} = W - Y$.

By the additional constraint, it is easy to see that the transitivity axiom holds in the semantics.

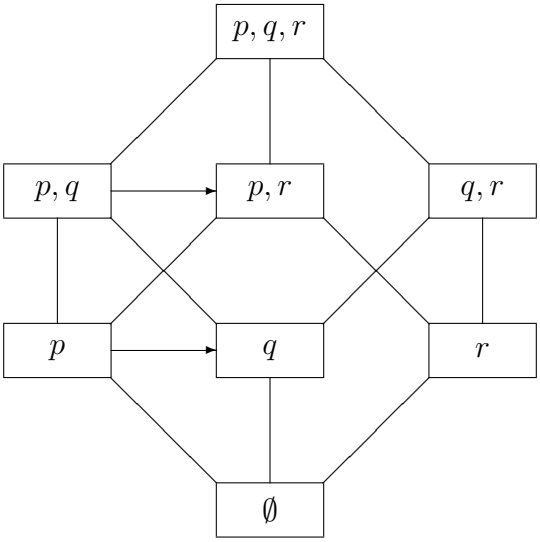


Figure 9.5: Counter Example Against the Transitivity

In ALX logics, we will use an improved MCP semantics, called *MCP*$^+$ *semantics*, to deal with the preference operator. For this semantics, we modify the old preference logic, i.e., which is based on the standard MCP semantics, with respect to the following points:

i) allowing preferences over preferences in the formal syntax.

ii) introducing a comparison relation on the powerset of W rather than on the possible world set themselves.

Formally, we define the comparison relation $\succ$ as follows:

$\succ \subseteq \mathcal{P}(W) \times \mathcal{P}(W)$, which satisfy the following conditions:

(NORM) $(\emptyset \not\succ X)$, $(X \not\succ \emptyset)$.
(TRAN) $cw(w, X \cap \overline{Y}) \succ cw(w, Y \cap \overline{X})$ and $cw(w, Y \cap \overline{Z}) \succ cw(w, Z \cap \overline{Y}) \Rightarrow cw(w, X \cap \overline{Z}) \succ cw(w, Z \cap \overline{X})$.
where $\overline{Y} = W - Y$.

In the next chapter, we will offer more details about the MCP$^+$ semantics, and prove its soundness and completeness.

Chapter 10

ALX1: A Propositional ALX Logic

10.1 Introduction

ALX1 is a propositional ALX logic, which combines an update logic, a preference logic, and a dynamic logic together. As argued in section 8.2, this combination is appropriate for an action logic for agents with bounded rationality. Many social theories are action-oriented. so a dynamic logic is needed. Secondly, preference plays an important role in social theories. So ALX1 contains a preference logic. Third, update offers a tool to formalize the counterfactual. Therefore, update logic is a sub-system of ALX1. A conditional logic also is an alternative which can play a similar role like an update logic. In the next chapter "ALX3: a multi-agent ALX logic", we will study the alternative in details.

10.2 Formal Syntax and Semantics

10.2.1 Syntax

ALX1 is a multimodal propositional logic. The propositional alphabet consists of a countable set of lower-case Latin symbols p_i to denote primitive propositions. The action alphabet has a finite set of actions a_i. Lower case Greek letters $\phi, \psi, \rho...$ (with or without subscript) denote well-formed formulae. We use $\langle a \rangle \psi$ to denote the one-place existential accessibility relation for action a, and $\boldsymbol{P}$ to denote the two-place preference relation. The symbol $\circ$ denotes a two-place operator for *updates*; updates are changes caused by an action. Note that updates in ALX1 refer to real state changes, not epistemological ones [Grahne 1991], so an update does not produce a new knowledge state, but a new situation.

10.2.1. Definition. (Syntax) *Let* $ATOM = \{p_i : i < \omega\}$, *and* $ACTION = \{a_1, ...a_k\}$ *for some* $k \in \omega$ *with* ω *standing for the ordinality of natural numbers. The set of formulae* FML *is defined recursively as follows:*

- $ATOM \subseteq FML$.

- $\phi \in FML \Rightarrow \neg\phi \in FML$.
- $\phi, \psi \in FML \Rightarrow (\phi \wedge \psi) \in FML$.
- $\phi \in FML, a \in ACTION \Rightarrow (\langle a \rangle \phi) \in FML$.
- $\phi, \psi \in FML \Rightarrow (\phi \circ \psi) \in FML$.
- $\phi, \psi \in FML \Rightarrow (\phi \boldsymbol{P} \psi) \in FML$.

Define $\bot$ as $\phi \wedge \neg\phi$ for an arbitrary ϕ, and $[a]\phi$ as $\neg\langle a \rangle \neg\phi$. Define the boolean connectives $\{\vee, \rightarrow, \leftrightarrow\}$, and the truth constant $\top$ from the given boolean connectives in the usual way.

10.2.2 Semantics

10.2.2. Definition. (ALX1 models) *We call* $M = \langle W, cw, \succ, \{R^a\}_{a \in ACTION}, V \rangle$ *an ALX1 model if*

- W *is a set of possible worlds,*
- $cw : W \times \mathcal{P}(\mathcal{W}) \hookrightarrow \mathcal{P}(\mathcal{W})$ *is a closest world function,*
- $\succ \subseteq \mathcal{P}(\mathcal{W}) \times \mathcal{P}(\mathcal{W})$ *is a comparison relation for preferences,*
- $R^a \subseteq W \times W$ *is an accessibility relation for each* a *in* $ACTION$*,*
- $V : ATOM \rightarrow \mathcal{P}(\mathcal{W})$ *is an assignment function for primitive propositions*

and (i) cw satisfies the conditions (CS1), (CS2), and (CSC);
and (ii) $\succ$ *satisfies the following conditions:*

(*NORM*) : $(\emptyset \not\succ X),\ (X \not\succ \emptyset)$.
(*TRAN*) : $cw(w, X \cap \overline{Y}) \succ cw(w, Y \cap \overline{X})$ *and* $cw(w, Y \cap \overline{Z}) \succ cw(w, Z \cap \overline{Y})$ $\Rightarrow cw(w, X \cap \overline{Z}) \succ cw(w, Z \cap \overline{X})$
where $\overline{Y} = W - Y$.

As argued in section 8.3.3, (CS1) through (CSC) constrain the closest-world function for the update. (CS1) ensures that the closest ϕ-worlds (relative to a given world) are indeed ϕ-worlds; (CS2) ensures that w is its own (and unique) closest ϕ-world if ϕ is true at w. (CSC) says that if ψ is true at the closest ϕ-world, then the closest ϕ-world is also a closest ϕ-and-ψ-world. (NORM) and (TRAN) constrain the semantic preference relation. They require normality and transitivity; "normality" stipulates that no comparison between two sets of worlds would involve empty set of worlds.

In the following, we use $M = \langle W, cw, \succ, R^a, V \rangle$ to denote $M = \langle W, cw, \succ, \{R^a\}_{a \in ACTION}, V \rangle$ if the omission cannot cause an ambiguity.

10.2.3. Definition. (Meaning function) *Let* FML *be as above and let:*

$$M = \langle W, cw, \succ, R^a, V \rangle$$

be an ALX1 model. The meaning function $[\![\ ]\!]_M : FML \rightarrow \mathcal{P}(\mathcal{W})$ *is defined as follows:*

$$
\begin{array}{lcl}
[\![p_i]\!]_M & = & V(p_i). \\
[\![\neg\phi]\!]_M & = & W \setminus [\![\phi]\!]_M. \\
[\![\phi \wedge \psi]\!]_M & = & [\![\phi]\!]_M \cap [\![\psi]\!]_M. \\
[\![\langle a\rangle\phi]\!]_M & = & \{w \in W : \exists w' \in W(R^a w w' \text{ and } w' \in [\![\phi]\!]_M)\}. \\
[\![\phi \circ \psi]\!]_M & = & \{w \in W : \exists w' \in W(w' \in [\![\phi]\!]_M \text{ and } w \in cw(w', [\![\psi]\!]_M))\}. \\
[\![\phi \boldsymbol{P} \psi]\!]_M & = & \{w \in W : cw(w, [\![\phi \wedge \neg\psi]\!]_M) \succ cw(w, [\![\neg\phi \wedge \psi]\!]_M)\}.
\end{array}
$$

The interpretation of the primitive propositions and the boolean connectives is straightforward. The interpretation of $\langle a\rangle\phi$ yields the set of worlds from where the agent can access at least one ϕ-worlds via action a. We use the "existential" version of the action modality, because real-life decisions typically depend on the *possibility* of a specific action in a specific situation.

Define the forcing relation as: [1]

$$M, w \| \!- \phi \stackrel{\text{def}}{\Longleftrightarrow} w \in [\![\phi]\!]_M.$$

10.2.4. Definition. (The logic ALX1) *Let FML be as above, let Mod be the class of all ALX1 models, and let $[\![\ ]\!]_M$ be as above too, defined for every model $M \in Mod$. We call the logic $ALX1 = \langle FML, Mod, [\![\ ]\!]_M\rangle$ ALX1 logic. $\models$ is defined as usual:*

$$M = \langle W, cw, \succ, R^a, V\rangle \models \phi \stackrel{\text{def}}{\Longleftrightarrow} (\forall w \in W)(M, w \| \!- \phi).$$

$$M \models \Gamma \stackrel{\text{def}}{\Longleftrightarrow} (\forall \gamma \in \Gamma)(M \models \gamma).$$

$$Mod(\Gamma) \stackrel{\text{def}}{\Longleftrightarrow} \{M \in Mod : M \models \Gamma\}.$$

$$\Gamma \models \phi \stackrel{\text{def}}{\Longleftrightarrow} Mod(\Gamma) \subseteq Mod(\{\phi\}).$$

Definitions 2-4 provide a semantic characterization of $ALX1$. The next definition provides a complete syntactic characterization.

Let $\vdash$ denote the notion of syntactic inference.

10.2.5. Definition. (ALX1 inference system) *Let ALX1S be the following set of*

[1] Use of the symbol $\| \!-$ in this chapter as an alternatives for $\models$; this is for histric reasons only.

axioms and rules of inference.

$$
\begin{array}{lll}
\textbf{(BA)}: & \text{all propositional tautologies.} & \\
 & & \\
\textbf{(A1)}: & \langle a\rangle\bot & \leftrightarrow \ \bot. \\
\textbf{(A2)}: & \langle a\rangle(\phi \vee \psi) & \leftrightarrow \ \langle a\rangle\phi \ \vee \langle a\rangle\psi. \\
 & & \\
\textbf{(U1)}: & \phi \circ \psi & \rightarrow \ \psi. \\
\textbf{(U2)}: & \phi \wedge \psi & \rightarrow \ \phi \circ \psi. \\
\textbf{(U3)}: & \neg(\phi \circ \bot), & \neg(\bot \circ \phi). \\
\textbf{(U4)}: & (\phi \vee \psi) \circ \chi & \leftrightarrow \ \phi \circ \chi \vee \psi \circ \chi. \\
\textbf{(U5)}: & (\phi \wedge \psi) \circ \psi & \rightarrow \ \phi. \\
\textbf{(U6)}: & (\phi \circ \psi) \wedge \chi & \rightarrow \ \phi \circ (\psi \wedge \chi). \\
 & & \\
\textbf{(CEP)}: & \phi \boldsymbol{P} \psi & \leftrightarrow \ (\phi \wedge \neg\psi)\boldsymbol{P}(\neg\phi \wedge \psi). \\
\textbf{(TR)}: & (\phi \boldsymbol{P} \psi) \wedge (\psi \boldsymbol{P} \chi) & \rightarrow \ (\phi \boldsymbol{P} \chi). \\
\textbf{(N)}: & \neg(\bot \boldsymbol{P} \phi), & \neg(\phi \boldsymbol{P} \bot). \\
 & & \\
\textbf{(MP)}: & \vdash \phi \ \& \ \vdash \phi \rightarrow \psi & \Rightarrow \ \vdash \psi. \\
\textbf{(NECA)}: & \vdash \phi & \Rightarrow \ \vdash [a]\phi. \\
\textbf{(MONA)}: & \vdash \langle a\rangle\phi \ \& \ \vdash \phi \rightarrow \psi & \Rightarrow \ \vdash \langle a\rangle\psi. \\
\textbf{(MONU)}: & \vdash \phi \circ \psi \ \& \ \vdash \phi \rightarrow \phi' & \Rightarrow \ \vdash \phi' \circ \psi. \\
\textbf{(SUBA)}: & \vdash (\phi \leftrightarrow \phi') & \Rightarrow \ \vdash (\langle a\rangle\phi) \leftrightarrow (\langle a\rangle\phi'). \\
\textbf{(SUBU)}: & \vdash (\phi \leftrightarrow \phi') \ \& \ \vdash (\psi \leftrightarrow \psi') & \Rightarrow \ \vdash (\phi \circ \psi) \leftrightarrow (\phi' \circ \psi'). \\
\textbf{(SUBP)}: & \vdash (\phi \leftrightarrow \phi') \ \& \ \vdash (\psi \leftrightarrow \psi') & \Rightarrow \ \vdash (\phi \boldsymbol{P} \psi) \leftrightarrow (\phi' \boldsymbol{P} \psi').
\end{array}
$$

Most axioms are straightforward. As usual, we have the propositional tautologies (BA). Since ALX1 is a *normal* modal logic, the absurdum is not true anywhere, so it is not accessible (A1). The action modalities behave as usual, so they distribute over disjunction both ways, but over conjunction only in one direction (A2). Indeed, we can get to ϕ-or-ψ-worlds via action a if and only if we can get via a to a ϕ-world or to a ψ-world. However, being able to get to ϕ-worlds via action a *and* being able to get to ψ-worlds via action a does not necessarily mean that a can get us to a world that is both ϕ-and-ψ.

As mentioned above, $\circ$ is a backward-looking operator. So, a successful ψ-update ends up in a ψ-world (U1), and the truth of both ϕ and ψ at a world allows us to perform a vacuous ψ-update, i.e., stay at that world (U2). (U3) reiterates the normality-condition for updates. Since there is no world where the absurdum is true, an update with the absurdum cannot succeed. (U4) expresses the left distribution of the disjunction over the update operator. The intuition is that if we've gotten to a χ-world from a ϕ- *or* a ψ-world, we've updated either from a ϕ-world or from a ψ-world. (U5) tells us that a void update is not going to change conditions. (U6) posits that if χ holds after updating ϕ with ψ, then we can update ϕ with $\psi \wedge \chi$ and obtain the same result. Reader more familiar with closest-world functions may already sense how the update operator will mimick the closest world function in

the syntax, helping to construct of a canonical model during the completeness proof. The axioms for the preference-operator expresses the conjunction expansion principle (CEP), transitivity (TR), and normality (N). So, if we prefer ϕ to ψ, we will also prefer the absence of ψ to the absence of ϕ. If we prefer ϕ to ψ, we are apt to prefer ϕ-and-not-ψ to ψ-and-not-ϕ. We have transitivity because we think that it is a natural principle of preference orders. We have irreflexivity because we are working with a strong preference relation, and we have normality to avoid inconsistent preference statements.

We have the modus ponens and the necessitation rule for the universal action modality (NECA), and monotonicity for the existential action modality. For the update-operator, we have *left* monotonicity, but not right monotonicity, the intuition being that a move from a ϕ world to the closest ψ-world w might end up at a different world than the move to the closest ψ'-world even if ψ implies ψ' at w. Logically equivalent propositions are substitutional in action- update- and preference formulae (SUBA), (SUBU), (SUBP). Note that we are *not* having monotonicity for preferences. Because of this, we are able to avoid the counterintuitive deductive closure of goals.

10.3 Formal Properties of ALX1

10.3.1. PROPOSITION. (Soundness of ALX1S) *The axioms (BA), (A1), (A2), (U1)-(U6), (CEP)-(N), and the inference rules (MP)-(SUBP) are sound for the class of ALX1 models.*

PROOF.
(U1) $\phi \circ \psi \rightarrow \psi$.

$$
\begin{array}{lll}
 & M, w \Vdash \phi \circ \psi & \\
\Leftrightarrow & \exists i \in [\![\phi]\!]_M (w \in cw(i, [\![\psi]\!]_M)) & \text{(Truth condition)} \\
\Rightarrow & \exists i \in [\![\phi]\!]_M (w \in [\![\psi]\!]_M) & \text{(CS1)} \\
\Rightarrow & w \in [\![\psi]\!]_M & \text{(Meta reasoning)} \\
\Leftrightarrow & M, w \Vdash \psi & \text{(Definition of } \Vdash\text{)}
\end{array}
$$

(U2) $\phi \wedge \psi \rightarrow \phi \circ \psi$.

$$
\begin{array}{lll}
 & M, w \Vdash \phi \wedge \psi & \\
\Leftrightarrow & M, w \Vdash \phi \text{ and } M, w \Vdash \psi & \text{(Truth condition)} \\
\Leftrightarrow & w \in [\![\phi]\!]_M \text{ and } w \in [\![\psi]\!]_M & \text{(Definition of } \Vdash\text{)} \\
\Rightarrow & w \in [\![\phi]\!]_M \text{ and } cw(w, [\![\psi]\!]_M) = \{w\} & \text{(CS2)} \\
\Rightarrow & \exists w (w \in [\![\phi]\!]_M \text{ and } w \in cw(w, [\![\psi]\!]_M)) & \text{(Meta reasoning)} \\
\Leftrightarrow & M, w \Vdash \phi \circ \psi & \text{(Truth condition)}
\end{array}
$$

A ψ-update results in ψ (U1), and the truth of both ϕ and ψ at a world that we could perform a vacuous ψ-update (and stay at that world) (U2).

(U3) $\neg(\bot \circ \phi)$.

$$\begin{array}{lll} & M, w\|\!\!-\bot \circ \phi & \\ \Leftrightarrow & \exists i(i \in [\![\bot]\!]_M \text{ and } w \in cw(i, [\![\phi]\!]_M)) & \text{(Truth condition)} \\ \Rightarrow & \exists i(i \in [\![\bot]\!]_M) & \text{(Meta reasoning)} \\ \Rightarrow & False & \text{(Meta reasoning)} \end{array}$$

Therefore, $\neg(\bot \circ \phi)$.

(U4) $(\phi \vee \psi) \circ \chi \leftrightarrow (\phi \circ \chi) \vee (\psi \circ \chi)$.

$$\begin{array}{lll} & M, w\|\!\!-(\phi \vee \psi) \circ \chi & \\ \Leftrightarrow & \exists i(i \in [\![\phi \vee \psi]\!]_M \text{ and } w \in cw(i, [\![\chi]\!]_M)) & \text{(Truth condition)} \\ \Leftrightarrow & \exists i((i \in [\![\phi]\!]_M \text{ and } w \in cw(i, [\![\chi]\!]_M)) \text{ or} & \\ & (i \in [\![\psi]\!] \text{ and } w \in cw(i, [\![\chi]\!]_M))) & \text{(Meta reasoning)} \\ \Leftrightarrow & w\|\!\!-(\phi \circ \chi) \text{ or } w\|\!\!-(\psi \circ \chi) & \text{(Truth condition)} \\ \Leftrightarrow & w\|\!\!-(\phi \circ \chi) \vee (\psi \circ \chi) & \text{(Truth condition)} \end{array}$$

(U5) $(\phi \wedge \psi) \circ \psi \rightarrow \phi$.

$$\begin{array}{lll} & M, w\|\!\!-(\phi \wedge \psi) \circ \psi & \\ \Leftrightarrow & \exists i(i \in [\![\phi \wedge \psi]\!]_M \text{ and } w \in cw(i, [\![\psi]\!]_M)) & \text{(Truth condition)} \\ \Rightarrow & \exists i(i \in [\![\phi]\!]_M \text{ and } i \in [\![\psi]\!]_M \text{ and } w \in cw(i, [\![\psi]\!]_M)) & \text{(Truth condition)} \\ \Rightarrow & \exists i(i \in [\![\phi]\!]_M \text{ and } w = i) & \text{(CS2)} \\ \Rightarrow & w \in [\![\phi]\!]_M & \text{(Meta reasoning)} \\ \Rightarrow & M, w\|\!\!-[\![\phi]\!]_M & \text{(Definition of } \|\!\!-) \end{array}$$

(U6) $(\phi \circ \psi) \wedge \chi \rightarrow \phi \circ (\psi \wedge \chi)$

$$\begin{array}{lll} & M, w\|\!\!-(\phi \circ \psi) \wedge \chi & \\ \Leftrightarrow & \exists i(i \in [\![\phi]\!]_M \text{ and } w \in cw(i, [\![\psi]\!]_M)) \text{ and } w \in [\![\chi]\!]_M & \text{(Truth condition)} \\ \Rightarrow & \exists i(i \in [\![\phi]\!]_M \text{ and } w \in cw(i, [\![\psi]\!]_M) \cap [\![\chi]\!]_M) & \text{(Meta reasoning)} \\ \Rightarrow & \exists i(i \in [\![\phi]\!]_M \text{ and } w \in cw(i, [\![\psi \wedge \chi]\!]_M)) & \text{(CSC)} \\ \Leftrightarrow & M, w\|\!\!-\phi \circ (\psi \wedge \chi) & \text{(Truth condition)} \end{array}$$

The proofs of the other axioms and rules are straightforward. □

10.3.2. Proposition. (More properties of update) *The following propositions are sound for the class of ALX1 models:*

$(U1^\circ)$ $\phi \circ \phi \rightarrow \phi$.
$(U2^\circ)$ $(\phi \circ \psi) \circ \psi) \rightarrow \phi \circ \psi$.
$(U3^\circ)$ $\neg\phi \wedge (\phi \circ \psi) \rightarrow (\phi \wedge \neg\psi) \circ \psi$.
$(U4^\circ)$ $(\phi \circ \psi \rightarrow \bot) \rightarrow ((\phi \rightarrow \bot) \vee (\psi \rightarrow \bot))$.
$(U5^\circ)$ $(\phi \circ \psi) \wedge \neg\phi \rightarrow \neg\psi \circ \psi$.

$(U6^\circ)$ $(\neg\phi \circ \psi) \wedge (\neg\psi \circ \phi) \rightarrow (\neg\phi \circ \phi) \wedge (\neg\psi \circ \psi)$.
$(U7^\circ)$ $(\phi \wedge \psi) \circ \phi \leftrightarrow (\phi \wedge \psi) \circ \psi$.
$(U8^\circ)$ $((\rho \wedge \phi) \circ \phi) \wedge \psi \rightarrow (\rho \wedge \psi) \circ \phi$.

PROOF.
$(U1^\circ)$ $\phi \circ \phi \rightarrow \phi$.

$$\begin{array}{lll} & M, w\Vdash\!\!-(\phi \circ \phi) & \\ \Leftrightarrow & \exists i(i \in [\![\phi]\!]_M \text{ and } w \in cw(i, [\![\phi]\!]_M)) & \text{(Truth condition)} \\ \Leftrightarrow & w \in [\![\phi]\!]_M & \text{(CS2)} \\ \Leftrightarrow & M, w\Vdash\!\!-\phi & \text{(Definition of } \Vdash\!\!-) \end{array}$$

$(U2^\circ)$ $(\phi \circ \psi) \circ \psi \leftrightarrow \phi \circ \psi$.

$(\Rightarrow)$

$$\begin{array}{lll} & M, w\Vdash\!\!-(\phi \circ \psi) \circ \psi & \\ \Leftrightarrow & \exists i(i \in [\![\phi \circ \psi]\!]_M \text{ and } w \in cw(i, [\![\psi]\!]_M)) & \text{(Truth condition)} \\ \Leftrightarrow & \exists i \exists j(j \in [\![\phi]\!]_M \text{ and } i \in cw(j, [\![\psi]\!]_M) \text{ and } w \in cw(i, [\![\psi]\!]_M)) & \text{(Truth condition)} \\ \Rightarrow & \exists i \exists j(j \in [\![\phi]\!]_M \text{ and } i \in cw(j, [\![\psi]\!]_M) \text{ and } w = i) & \text{(CS2)} \\ \Leftrightarrow & \exists j(j \in [\![\phi]\!]_M \text{ and } w \in cw(j, [\![\psi]\!]_M)) & \text{(Meta reasoning)} \\ \Leftrightarrow & M, w\Vdash\!\!-\phi \circ \psi & \text{(Truth condition)} \end{array}$$

$(\Leftarrow)$

$$\begin{array}{lll} & M, w\Vdash\!\!-(\phi \circ \psi) & \\ \Rightarrow & M, w\Vdash\!\!-(\phi \circ \psi) \wedge \psi & \text{(U1)} \\ \Rightarrow & M, w\Vdash\!\!-(\phi \circ \psi) \circ \psi & \text{(U2)} \end{array}$$

Therefore, $(\phi \circ \psi) \leftrightarrow (\phi \circ \psi) \circ \psi$.

$(U3^\circ)$ $\neg\phi \wedge (\phi \circ \psi) \rightarrow (\phi \wedge \neg\psi) \circ \psi$.

$$\begin{array}{lll} & M, w\Vdash\!\!-\neg\phi \wedge (\phi \circ \psi) & \\ \Leftrightarrow & M, w\Vdash\!\!-\neg\phi \text{ and } M, w\Vdash\!\!-\phi \circ \psi & \text{(Truth condition)} \\ \Leftrightarrow & M, w\Vdash\!\!-\neg\phi \text{ and } \exists i((i \in [\![\phi]\!]_M) \text{ and } w \in cw(i, [\![\psi]\!]_M)) & \text{(Truth condition)} \end{array}$$

Case 1: $i \in [\![\neg\psi]\!]_M$,

$$\begin{array}{lll} & i \in [\![\neg\psi]\!]_M & \\ \Rightarrow & \exists i(i \in [\![\phi]\!]_M \text{ and } i \in [\![\neg\psi]\!]_M \text{ and } w \in cw(i, [\![\psi]\!]_M)) & \text{(Assumption)} \\ \Leftrightarrow & \exists i(i \in [\![\phi \wedge \neg\psi]\!]_M \text{ and } w \in cw(i, [\![\psi]\!]_M)) & \text{(Truth condition)} \\ \Leftrightarrow & M, w\Vdash\!\!-(\phi \wedge \neg\psi) \circ \psi & \text{(Truth condition)} \end{array}$$

Case 2: $i \in [\![\psi]\!]_M$,

$$
\begin{array}{lll}
& i \in [\![\psi]\!]_M & \\
\Rightarrow & i = w & (w \in cw(i, [\![\psi]\!]_M), \text{ and (CS2)}) \\
\Rightarrow & M, w \Vdash \phi \text{ and } M, w \Vdash \neg\phi & (i \in [\![\phi]\!]_M, \text{ and } M, w \Vdash \neg\phi) \\
\Rightarrow & \textbf{False} &
\end{array}
$$

$(U4^\circ)$ $(\phi \circ \psi \rightarrow \bot) \rightarrow ((\phi \rightarrow \bot) \vee (\psi \rightarrow \bot))$.

$$
\begin{array}{ll}
& (\phi \wedge \psi) \rightarrow (\phi \circ \psi) \text{ is valid} \\
\Leftrightarrow & \neg(\phi \circ \psi) \rightarrow \neg\phi \vee \neg\psi \text{ is valid} \\
\Leftrightarrow & (\neg(\phi \circ \psi) \vee \bot) \rightarrow (\neg\phi \vee \bot) \vee (\neg\psi \vee \bot) \text{ is valid} \\
\Leftrightarrow & (\phi \circ \psi \rightarrow \bot) \rightarrow (\phi \rightarrow \bot) \vee (\psi \rightarrow \bot) \text{ is valid}
\end{array}
$$

$(U5^\circ)$ $(\phi \circ \psi) \wedge \neg\phi \rightarrow \neg\psi \circ \psi$.

$$
\begin{array}{lll}
& M, w \Vdash (\phi \circ \psi) \wedge \neg\phi & \\
\Leftrightarrow & \exists i(i \in [\![\phi]\!]_M \text{ and } w \in cw(i, [\![\psi]\!]_M) \text{ and } w \in [\![\neg\phi]\!]_M) & \text{(Truth condition)} \\
\Rightarrow & \exists i(i \neq w) & \text{(Meta reasoning)}
\end{array}
$$

Suppose that $M, i \Vdash \psi$, then
$M, i \Vdash \psi \Rightarrow cw(i, [\![\psi]\!]_M) = \{i\} \Rightarrow w = i \Rightarrow \textbf{False}$.

Therefore, $M, i \nVdash \psi$, namely, $M, w \Vdash \neg\psi \circ \psi$

$(U6^\circ)$ $(\neg\phi \circ \psi) \wedge (\neg\psi \circ \phi) \rightarrow (\neg\phi \circ \phi) \wedge (\neg\psi \circ \psi)$.

$$
\begin{array}{lll}
& M, w \Vdash (\neg\psi \circ \phi) \wedge (\neg\phi \circ \psi) & \\
\Rightarrow & M, w \Vdash (\neg\psi \circ \phi) \wedge \psi \wedge (\neg\phi \circ \psi) \wedge \phi & \text{(U1)} \\
\Rightarrow & M, w \Vdash (\neg\psi \circ \psi) \wedge (\neg\phi \circ \phi) & (U5^\circ)
\end{array}
$$

$(U7^\circ)$ $(\phi \wedge \psi) \circ \phi \leftrightarrow (\phi \wedge \psi) \circ \psi$.

$$
\begin{array}{lll}
& M, w \Vdash (\phi \wedge \psi) \circ \phi & \\
\Rightarrow & M, w \Vdash ((\phi \wedge \psi) \circ \phi) \wedge \phi & \text{(U1)} \\
\Rightarrow & M, w \Vdash (\psi \wedge \phi) & \text{(U5)} \\
\Rightarrow & M, w \Vdash (\psi \wedge \phi) \circ \psi & \text{(U2)}
\end{array}
$$

Therefore, $(\phi \wedge \psi) \circ \phi \rightarrow (\phi \wedge \psi) \circ \psi$.
The proof for $(\phi \wedge \psi) \circ \psi \rightarrow (\phi \wedge \psi) \circ \phi$ goes analogously.

$(U8^\circ)$ $((\rho \wedge \phi) \circ \phi) \wedge \psi \rightarrow (\rho \wedge \psi) \circ \phi$.

$$
\begin{array}{lll}
& M, w \Vdash ((\rho \wedge \phi) \circ \phi) \wedge \psi & \\
\Rightarrow & M, w \Vdash \rho \wedge ((\rho \wedge \phi) \circ \phi) \wedge \psi & \text{(U5)} \\
\Rightarrow & M, w \Vdash \rho \wedge \phi \wedge \psi & \text{(U1)} \\
\Rightarrow & M, w \Vdash (\rho \wedge \psi) \circ \phi & \text{(U2)}
\end{array}
$$

□

10.3.3. Proposition. (More properties of preference)
(CP) $\phi \boldsymbol{P} \psi \leftrightarrow (\neg\psi)\boldsymbol{P}(\neg\phi)$.
(IR) $\neg(\phi \boldsymbol{P} \phi)$.
(NT) $\neg(\top \boldsymbol{P} \phi)$, $\neg(\phi \boldsymbol{P} \top)$.
(AS) $\phi \boldsymbol{P} \psi \rightarrow \neg(\psi \boldsymbol{P} \phi)$.

PROOF. In the last chapter, we have already proved that (CEP) and (N) together imply (CP), (IR), and (NT). Therefore, they also hold in ALX1.

(AS) $\phi \boldsymbol{P} \psi \rightarrow \neg(\psi \boldsymbol{P} \phi)$.

$$\begin{array}{lll} & \vdash (\phi \boldsymbol{P} \psi) \wedge (\psi \boldsymbol{P} \phi) & \\ \Rightarrow & \vdash \phi \boldsymbol{P} \phi & \text{(TR)} \\ \Rightarrow & \vdash \bot & \text{(IR)} \end{array}$$

□

10.4 Completeness

The completeness proof for ALX1S proceeds along the lines of a Henkin-style construction. We give a detailed proof. So our task is to construct a canonical model that is an ALX1 model. First, we need two lemmas (for the action and the update operators, respectively) that ensure the existence of certain maximal consistent sets required in the construction of the canonical model. Let W_c be the set of all maximal consistent sets built from the elements of FML.

10.4.1. Lemma. (Action lemma) $\forall w \in W_c(\langle a \rangle \phi \in w \Rightarrow (\exists z \in W_c)(\phi \in z \textit{ and } (\forall \psi \in z)(\langle a \rangle \psi \in w)))$.

PROOF. Suppose that $\langle a \rangle \phi \in w$, and let $F = \{\phi\} \cup \{\psi : \neg\langle a \rangle \neg\psi \in w\}$. Let $w^* = \{\psi : \neg\langle a \rangle \neg\psi \in w\}$.

We show first that (1) w^*, and (2) F are consistent. We then show (3) that we can always extend F to an F' such that F' satisfies the condition of the lemma, i.e, $F' = z$.

(1) we claim that w^* is consistent. This is implied by

(1.1) Assume that $\bot \in w^*$ we then show that $\bot \in w^* \Rightarrow$ **False**.

$$
\begin{array}{lll}
 & \bot \in w^* & \\
\Rightarrow & \neg\langle a\rangle\neg\bot \in w & \text{(Definition of } w^*\text{)} \\
\Rightarrow & \neg\langle a\rangle\top \in w & \text{(Propositional logic)} \\
\Rightarrow & \neg\langle a\rangle\top \wedge \langle a\rangle\phi \in w & \text{(Assumption)} \\
\Rightarrow & \neg\langle a\rangle\top \wedge \langle a\rangle\top \in w & \text{(MONA)} \\
\Rightarrow & \textbf{False} & \text{(Maximal consistency of } w\text{)}
\end{array}
$$

(1.2) We show that $\phi_1, \phi_2 \in w^* \Rightarrow (\phi_1 \wedge \phi_2) \in w^*$.

$$
\begin{array}{lll}
 & \phi_1, \phi_2 \in w^* & \\
\Rightarrow & \neg\langle a\rangle\neg\phi_1 \in w \text{ and } \neg\langle a\rangle\neg\phi_2 \in w & \text{(Definition of } w^*\text{)} \\
\Rightarrow & \neg(\langle a\rangle\neg\phi_1 \vee \langle a\rangle\neg\phi_2) \in w & \text{(Propositional Logic)} \\
\Rightarrow & \neg(\langle a\rangle(\neg\phi_1 \vee \neg\phi_2)) \in w & \text{(A2)} \\
\Rightarrow & \neg(\langle a\rangle\neg(\phi_1 \wedge \phi_2)) \in w & \text{(Propositional logic)} \\
\Rightarrow & \phi_1 \wedge \phi_2 \in w^* & \text{(Definition of } w^*\text{)}
\end{array}
$$

(1.3) For arbitrary ψ, we must show that $\psi \in w^*$,$\neg\psi \in w^* \Rightarrow$ **False**.

$$
\begin{array}{lll}
 & \psi \in w^*, \neg\psi \in w^* & \\
\Rightarrow & (\psi \wedge \neg\psi) \in w^* & (1.2) \\
\Rightarrow & \bot \in w^* & \text{(Propositional Logic)} \\
\Rightarrow & \textbf{False} & (1.1)
\end{array}
$$

We conclude that w^* is consistent.

(2) We claim that F is consistent. This is implied by

(2.1) $\vdash \phi \rightarrow \bot \Rightarrow$ **False**.
(2.2) For any $\psi \in w^*$,$\vdash (\phi \wedge \psi \rightarrow \bot) \Rightarrow$ **False**.

(2.1) Assume that $\vdash \phi \rightarrow \bot$, then,

$$
\begin{array}{lll}
 & \vdash \phi \rightarrow \bot & \\
\Rightarrow & \langle a\rangle\bot \in w & (\langle a\rangle\phi \in w \text{ and (MONA)}) \\
\Rightarrow & \bot \in w & \text{(A1)} \\
\Rightarrow & \textbf{False} & \text{(Maximal consistency of } w\text{)}
\end{array}
$$

(2.2) For arbitrary $\psi \in w^*$, assume that $\vdash (\phi \wedge \psi \rightarrow \bot)$. We show that $\vdash (\phi \wedge \psi \rightarrow \bot) \Rightarrow$ **False**.

$$
\begin{array}{lll}
 & \vdash (\phi \wedge \psi \rightarrow \bot) & \\
\Rightarrow & \vdash (\phi \rightarrow \neg\psi) & \text{(Propositional logic)} \\
\Rightarrow & \vdash (\phi \rightarrow \neg\psi) \text{ and } \psi \in w^* \text{ and } \langle a\rangle\phi \in w & \text{(Assumption)} \\
\Rightarrow & \vdash (\phi \rightarrow \neg\psi) \text{ and } \neg\langle a\rangle\neg\psi \in w \text{ and } \langle a\rangle\phi \in w & \text{(Definition of } w^*\text{)} \\
\Rightarrow & \neg\langle a\rangle\neg\psi \in w \text{ and } \langle a\rangle\neg\psi \in w & \text{(MONA)} \\
\Rightarrow & \textbf{False} & \text{(Maximal consistency of } w\text{)}
\end{array}
$$

We conclude that F is consistent. We show now that any maximal extension F' of F satisfies the lemma. So, let F' be an arbitrary maximal consistent extension of F. We must show that

(3.1) F' exists.
(3.2) $\phi \in F'$,
(3.3) $\langle a \rangle \psi \notin w \Rightarrow \psi \notin F'$.

(3.1) Straightforward from Lindenbaum's lemma.

(3.2) From the definition of F'.

(3.3) We have:

	$\langle a \rangle \psi \notin w$	
$\Rightarrow$	$\neg \langle a \rangle \psi \in w$	(Maximal consistency of w)
$\Rightarrow$	$\neg \langle a \rangle \neg \neg \psi \in w$	(Propositional Logic)
$\Rightarrow$	$\neg \psi \in w^*$	(Definition of w^*)
$\Rightarrow$	$\neg \psi \in F$	(Definition of F)
$\Rightarrow$	$\neg \psi \in F'$	(Definition of F')
$\Rightarrow$	$\psi \notin F'$	(Maximal consistency of F')

□

The next lemma parallels the action lemma for the update operator.

10.4.2. Lemma. (Update lemma) $\forall w \in W_c(\phi \circ \chi \in w \Rightarrow (\exists z \in W_c)(\phi \in z \text{ and } (\forall \psi \in z)(\psi \circ \chi \in w)))$.

PROOF. Suppose that $\phi \circ \chi \in w$, and let $F = \{\phi\} \cup \{\psi : \neg(\neg \psi \circ \chi) \in w\}$. Let $w^\circ = \{\psi : \neg(\neg \psi \circ \chi) \in w\}$. The proof's structure parallels the action lemma. We show first that (1) w°, and (2) F are consistent. We then show (3) that we can always extend F to an F' such that F' satisfies the condition of the lemma, i.e, $F' = z$.

(1) We claim that w° is consistent.
(1.1) Assume that $\bot \in w^\circ$, then we can show that $\bot \in w^\circ \Rightarrow$ **False**.

	$\bot \in w^\circ$	
$\Rightarrow$	$\neg(\neg \bot \circ \chi) \in w$	(Definition of w°)
$\Rightarrow$	$\neg(\top \circ \chi) \in w$	(Propositional logic)
$\Rightarrow$	$\neg(\top \circ \chi) \wedge (\phi \circ \chi) \in w$	(Assumption)
$\Rightarrow$	$\neg(\top \circ \chi) \wedge (\top \circ \chi) \in w$	(MONU)
$\Rightarrow$	**False**	(Maximal consistency of w)

(1.2) We show that $\phi_1, \phi_2 \in w^\circ \Rightarrow (\phi_1 \wedge \phi_2) \in w^\circ$.

	$\phi_1, \phi_2 \in w^\circ$	
$\Rightarrow$	$\neg(\neg\phi_1 \circ \chi) \in w$ and $\neg(\neg\phi_2 \circ \chi) \in w$	(Definition of w°)
$\Rightarrow$	$\neg(\neg\phi_1 \circ \chi \vee \neg\phi_2 \circ \chi) \in w$	(Maximal consistency of w)
$\Rightarrow$	$\neg((\neg\phi_1 \vee \neg\phi_2) \circ \chi) \in w$	(U4)
$\Rightarrow$	$\neg(\neg(\phi_1 \wedge \phi_2) \circ \chi) \in w$	(Propositional logic)
$\Rightarrow$	$\phi_1 \wedge \phi_2 \in w^\circ$	(Definition of w°)

(1.3) For arbitrary ψ, we must show that $\psi \in w^\circ$,$\neg\psi \in w^\circ \Rightarrow$ **False**.

	$\psi \in w^\circ, \neg\psi \in w^\circ$	
$\Rightarrow$	$(\psi \wedge \neg\psi) \in w^\circ$	(1.2)
$\Rightarrow$	$\bot \in w^\circ$	(Propositional Logic)
$\Rightarrow$	**False**	(1.1)

We conclude that w° is consistent.

(2) We claim that F is consistent. This is implied by

(2.1) $\vdash (\phi \rightarrow \bot) \Rightarrow$ **False**.
(2.2) For any $\psi \in w^\circ$,$\vdash (\phi \wedge \psi \rightarrow \bot) \Rightarrow$ **False**.

(2.1) Assume that $\vdash (\phi \rightarrow \bot)$, then,

	$\vdash (\phi \rightarrow \bot)$	
$\Rightarrow$	$\bot \circ \chi \in w$	($\phi \circ \chi \in w$ and (MONA))
$\Rightarrow$	$\bot \in w$	(U3)
$\Rightarrow$	**False**	(Maximal consistency of w)

(2.2) For arbitrary $\psi \in w^\circ$, assume that $\vdash (\phi \wedge \psi \rightarrow \bot)$. We show that $\vdash (\phi \wedge \psi \rightarrow \bot) \Rightarrow$ **False**.

	$\vdash (\phi \wedge \psi \rightarrow \bot)$	
$\Rightarrow$	$\vdash (\phi \rightarrow \neg\psi)$	(Propositional logic)
$\Rightarrow$	$\vdash (\phi \rightarrow \neg\psi)$ and $\psi \in w^\circ$ and $\phi \circ \chi \in w$	(Assumption)
$\Rightarrow$	$\vdash (\phi \rightarrow \neg\psi)$ and $\neg(\neg\psi \circ \chi) \in w$ and $\phi \circ \chi \in w$	(Definition of w°)
$\Rightarrow$	$\neg(\neg\psi \circ \chi) \in w$ and $(\neg\psi \circ \chi) \in w$	(MONU)
$\Rightarrow$	**False**	(Maximal consistency of w)

We conclude that F is consistent. We now show that any arbitrary maximal extension F' of F satisfies the lemma. So, let F' be an arbitrary maximal consistent extensions of F. We must show that

(3.1) F' exists.
(3.2) $\phi \in F'$
(3.3) $\psi \circ \chi \notin w \Rightarrow \psi \notin F'$

(3.1) Straightforward from Lindenbaum's lemma.

(3.2) From the definition of F'.

(3.3) We show that as follows:

$$
\begin{array}{lll}
 & \psi \circ \chi \notin w & \\
\Rightarrow & \neg(\psi \circ \chi) \in w & \text{(Maximal consistency of } w\text{)} \\
\Rightarrow & \neg(\neg\neg\psi \circ \chi) \in w & \text{(SUBU)} \\
\Rightarrow & \neg\psi \in w^{\circ} & \text{(Definition of } w^{\circ}\text{)} \\
\Rightarrow & \neg\psi \in F & \text{(Definition of } F\text{)} \\
\Rightarrow & \neg\psi \in F' & \text{(Definition of } F'\text{)} \\
\Rightarrow & \psi \notin F' & \text{(Maximal consistency of } F'\text{)}
\end{array}
$$

□

10.4.1. Theorem. (Completeness of ALXS) *ALX1S is complete for the class of ALX1 models.*

PROOF. We construct a canonical model $M_c = \langle W_c, cw, R^a, \succ, V\rangle$ and show that:

(1) $\chi \in w \in W_c \Leftrightarrow w \in [\![\chi]\!]_{M_c}$. (Truth Lemma)
(2) M_c is an ALX1 model.

Define $M_c = \langle W_c, cw, R^a, \succ, \rangle$ as follows:

$W_c = \{i \ : \ i \text{ is a maximal consistent set}\}$.

$w \in cw(j, [\![\psi]\!]_{M_c})$ iff $\forall\rho(\rho \in j \Rightarrow \rho \circ \psi \in w)$.

$\langle w, x\rangle \in R^a$ iff $\forall\rho(\rho \in x \Rightarrow \ \langle a\rangle\rho \in w)$.

$cw(w, [\![\phi \wedge \neg\psi]\!]_{M_c}) \succ cw(w, [\![\neg\phi \wedge \psi]\!]_{M_c})$ iff $\phi \boldsymbol{P} \psi \in w$.

$V(p_i) = \{w \ : \ p_i \in w\}$.

We prove the truth lemma by induction on the complexity of χ.

(1) $\chi \in w \Leftrightarrow w \in [\![\chi]\!]_{M_c}$.

The cases (1.1) $\chi \equiv p_i$, (1.2) $\chi \equiv \neg\phi$, and (1.3) $\chi \equiv \phi \wedge \psi$

are straightforward.

(1.4) $\chi \equiv \langle a \rangle \phi$.

$$
\begin{array}{lll}
 & \langle a \rangle \phi \in w & \\
\Rightarrow & \exists z \in W_c(\phi \in z \text{ and } \forall \psi \in z(\langle a \rangle \psi \in w)) & \text{(Action lemma)} \\
\Rightarrow & \exists z(\phi \in z \text{ and } R^a wz) & \text{(Definition of } R^a\text{)} \\
\Rightarrow & \exists z(z \in [\![\phi]\!]_{M_c} \text{ and } R^a wz) & \text{(Induction hypothesis)} \\
\Rightarrow & w \in [\![\langle a \rangle \phi]\!]_{M_c} & \text{(Truth condition)}
\end{array}
$$

$$
\begin{array}{lll}
 & w \in [\![\langle a \rangle \phi]\!]_{M_c} & \\
\Leftrightarrow & \exists z \in W_c(R^a wz \text{ and } z \in [\![\phi]\!]_{M_c}) & \text{(Truth condition)} \\
\Leftrightarrow & \exists z \in W_c(R^a wz \text{ and } \phi \in z) & \text{(Induction hypothesis)} \\
\Rightarrow & \langle a \rangle \phi \in w & \text{(Definition of } R^a\text{)}
\end{array}
$$

(1.5) $\chi \equiv \phi \circ \psi$.

$$
\begin{array}{lll}
 & \phi \circ \psi \in w & \\
\Rightarrow & \exists z(\phi \in z \text{ and } (\forall \rho \in z)((\rho \circ \psi) \in w)) & \text{(Update lemma)} \\
\Rightarrow & \exists z(\phi \in z \text{ and } w \in cw(z, [\![\psi]\!]_{M_c})) & \text{(Definition of } cw\text{)} \\
\Rightarrow & \exists z(z \in [\![\phi]\!]_{M_c} \text{ and } w \in cw(z, [\![\psi]\!]_{M_c})) & \text{(Induction hypothesis)} \\
\Rightarrow & w \in [\![\phi \circ \psi]\!]_{M_c} & \text{(Truth condition)}
\end{array}
$$

$$
\begin{array}{lll}
 & w \in [\![\phi \circ \psi]\!]_{M_c} & \\
\Leftrightarrow & \exists z(z \in [\![\phi]\!]_{M_c} \text{ and } w \in cw(z, [\![\psi]\!]_{M_c})) & \text{(Truth condition)} \\
\Leftrightarrow & \exists z(\phi \in z \text{ and } w \in cw(z, [\![\psi]\!]_{M_c})) & \text{(Induction hypothesis)} \\
\Rightarrow & \phi \circ \psi \in w & \text{(Definition of } cw\text{)}
\end{array}
$$

(1.6) $\chi \equiv \phi \boldsymbol{P} \psi$.

$$
\begin{array}{lll}
 & \phi \boldsymbol{P} \psi \in w & \\
\Leftrightarrow & cw(w, [\![\phi \wedge \neg\psi]\!]_{M_c}) \succ cw(w, [\![\psi \wedge \neg\phi]\!]_{M_c}) & \text{(Definition of } \succ\text{)} \\
\Leftrightarrow & w \in [\![\phi \boldsymbol{P} \psi]\!]_{M_c} & \text{(Truth condition)}
\end{array}
$$

This concludes the proof of the truth lemma. We now show that M_c is an ALX1 model. So, we have to show that cw satisfies (CS1), (CS2), and (CSC). Moreover, we have to show that $\succ$ satisfies the normality and transitivity conditions.

(CS1) $w \in cw(j, [\![\psi]\!]_{M_c}) \Rightarrow w \in [\![\psi]\!]_{M_c}$.

$$
\begin{array}{lll}
 & w \in cw(j, [\![\psi]\!]_{M_c}) & \\
\Leftrightarrow & \forall \rho(\rho \in j \Rightarrow \rho \circ \psi \in w) & \text{(Definition of } cw\text{)} \\
\Rightarrow & \exists \rho(\rho \in j \text{ and } \rho \circ \psi \in w) & (j \text{ is not an empty set}) \\
\Rightarrow & \psi \in w & \text{(U1)} \\
\Leftrightarrow & w \in [\![\psi]\!]_{M_c} & \text{(Truth lemma)}
\end{array}
$$

(CS2) $j \in [\![\psi]\!]_{M_c} \Rightarrow cw(j, [\![\psi]\!]_{M_c}) = \{j\}$.

We must show that:
(a) $j \in [\![\psi]\!]_{M_c} \Rightarrow j \in cw(j, [\![\psi]\!]_{M_c})$.
(b) $j \in [\![\psi]\!]_{M_c}$ and $j' \in cw(j, [\![\psi]\!]_{M_c}) \Rightarrow j = j'$.

For (a), we have:

	$j \in [\![\psi]\!]_{M_c}$	
$\Leftrightarrow$	$\psi \in j$	(Truth lemma)
$\Rightarrow$	$\forall\rho(\rho \in j \Rightarrow (\rho \wedge \psi) \in j)$	(Maximal consistency of j)
$\Rightarrow$	$\forall\rho(\rho \in j \Rightarrow (\rho \circ \psi) \in j)$	(U2)
$\Rightarrow$	$j \in cw(j, [\![\psi]\!]_{M_c})$	(Definition of cw)

For (b), suppose that $j \in [\![\psi]\!]_{M_c}$ and $j' \in cw(j, [\![\psi]\!]_{M_c})$, we first show that $j \subseteq j'$. Then by the maximal consistency of both j and j', we have $j = j'$. To show that $j \subseteq j'$, we proceed by reductio ad absurdum and show that $\rho \in j$ and $\rho \notin j' \Rightarrow$ **False** for arbitrary ρ.

	$\rho \in j$ and $\rho \notin j'$	
$\Leftrightarrow$	$\rho \in j$ and $\neg\rho \in j'$	(Maximal consistency of j')
$\Rightarrow$	$\rho \wedge \psi \in j$ and $\neg\rho \in j'$	($j \in [\![\psi]\!]_{M_c}$)
$\Rightarrow$	$((\rho \wedge \psi) \circ \psi) \in j'$ and $\neg\rho \in j'$	($j' \in cw(j, [\![\psi]\!]_{M_c})$)
$\Rightarrow$	$((\rho \wedge \psi) \circ \psi) \in j'$ and $\neg((\rho \wedge \psi) \circ \psi) \in j'$	(U5)
$\Rightarrow$	**False**	(Maximal consistency of j')

(CSC) $cw(w, [\![\phi]\!]_{M_c}) \cap [\![\psi]\!]_{M_c} \subseteq cw(w, [\![\phi \wedge \psi]\!]_{M_c})$.

For any $j \in cw(w, [\![\phi]\!]_{M_c}) \cap [\![\psi]\!]_{M_c}$, we have to show that $j \in cw(w, [\![\phi \wedge \psi]\!]_{M_c})$. That is, for any ρ, if $\rho \in w$, then $\rho \circ (\phi \wedge \psi) \in j$ by the definition of cw.

For any ρ,

	$\rho \in w$ and $j \in cw(w, [\![\phi]\!]_{M_c}) \cap [\![\psi]\!]_{M_c}$	
$\Rightarrow$	$\rho \in w$ and $j \in cw(w, [\![\phi]\!]_{M_c})$ and $\psi \in j$	(Truth lemma)
$\Rightarrow$	$\rho \circ \phi \in j$ and $\psi \in j$	(Definition of cw)
$\Rightarrow$	$(\rho \circ \phi) \wedge \psi \in j$	(Consistency of w)
$\Rightarrow$	$\rho \circ (\phi \wedge \psi) \in j$	(U6)

Therefore, $j \in cw(w, [\![\phi \wedge \psi]\!]_{M_c})$ by the definition of cw, so (CSC) holds.

(NORM) ($\emptyset \not\succ X$).

We must show that $\emptyset \succ X \Rightarrow$ **False**

$$
\begin{array}{lll}
 & \emptyset \succ X & \\
\Rightarrow & \exists w \exists \phi \exists \psi (\phi \boldsymbol{P} \psi \in w \text{ and } cw(w, [\![\phi \wedge \neg \psi]\!]_{M_c}) = \emptyset \text{ and} & \\
 & cw(w, [\![\psi \wedge \neg \phi]\!]_{M_c}) = X \text{ and} & \\
 & cw(w, [\![\phi \wedge \neg \psi]\!]_{M_c}) \succ cw(w, [\![\psi \wedge \neg \phi]\!]_{M_c})) & \text{(Definition of } \succ) \\
\Rightarrow & cw(w, [\![\bot]\!]_{M_c}) \succ cw(w, [\![\psi \wedge \neg \phi]\!]_{M_c}) & (cw(w, [\![\bot]\!]_{M_c}) = \emptyset) \\
\Rightarrow & cw(w, [\![\bot \wedge \neg(\psi \wedge \neg \phi)]\!]_{M_c}) \succ cw(w, [\![(\psi \wedge \neg \phi) \wedge \neg \bot]\!]_{M_c}) & \text{(Meta reasoning)} \\
\Rightarrow & \bot \boldsymbol{P} (\psi \wedge \neg \phi) \in w & \text{(Definition of } cw) \\
\Rightarrow & \textbf{False} & \text{(N)}
\end{array}
$$

(TRAN) $cw(w, X \cap \overline{Y}) \succ cw(w, Y \cap \overline{X})$ and
$cw(w, Y \cap \overline{Z}) \succ cw(w, Z \cap \overline{Y}) \Rightarrow cw(w, X \cap \overline{Z}) \succ cw(w, Z \cap \overline{X})$.

$$
\begin{array}{lll}
 & cw(w, X \cap \overline{Y}) \succ cw(w, Y \cap \overline{X}) \text{ and} & \\
 & cw(w, Y \cap \overline{Z}) \succ cw(w, Z \cap \overline{Y}) & \\
\Rightarrow & \exists \phi \exists \psi \exists \chi (\phi \boldsymbol{P} \psi \in w \text{ and } \psi \boldsymbol{P} \chi \in w \text{ and } [\![\phi]\!]_{M_c} = X \text{ and} & \\
 & [\![\psi]\!]_{M_c} = Y \text{ and } [\![\chi]\!]_{M_c} = Z) & \text{(Definition of } \succ) \\
\Rightarrow & \exists \phi \exists \psi \exists \chi (\phi \boldsymbol{P} \chi \in w \text{ and } \psi \boldsymbol{P} \chi \in w \text{ and } [\![\phi]\!]_{M_c} = X \text{ and} & \\
 & [\![\psi]\!]_{M_c} = Y \text{ and } [\![\chi]\!]_{M_c} = Z) & \text{(TR)} \\
\Rightarrow & \exists \phi \exists \psi \exists \chi (cw(w, [\![\phi \wedge \neg \chi]\!]_{M_c}) \succ cw(w, [\![\chi \wedge \neg \phi]\!]_{M_c}) \text{ and} & \\
 & [\![\phi]\!]_{M_c} = X \text{ and } [\![\psi]\!]_{M_c} = Y \text{ and } [\![\chi]\!]_{M_c} = Z) & \text{(Definition of } \succ) \\
\Rightarrow & cw(w, X \cap \overline{Z}) \succ cw(w, Z \cap \overline{X}) & \text{(Meta reasoning)}
\end{array}
$$

This concludes the proof that M_c is an ALX1 model. □

10.5 The Finite Model Property of ALX1

Naturally, we would like ALX1 to be decidable. As we will prove in this section, ALX1 has the finite model property, i.e., for each non-theorem ψ, there exists a finite model that provides a counterexample for ψ. Since ALX1 is recursively axiomatizable, ALX1 is decidable.

10.5.1. Definition. (Finite model property) *A logic S is said to have the finite model property (f.m.p.) iff, for arbitrary ϕ such that $\nvdash_S \phi$, there exists a finite model M such that*
(1) $\exists w (M, w \Vdash \neg \phi)$
(2) $\forall \rho (\vdash_S \rho \Rightarrow \forall w (M, w \Vdash \rho))$

10.5.2. Definition. (Subformula set) *A formula set Φ_ρ is said to be the subformula set of ρ iff Φ_ρ satisfies the following conditions:*

- $\rho \in \Phi_\rho$.
- $\neg \phi \in \Phi_\rho \Rightarrow \phi \in \Phi_\rho$.
- $\phi \wedge \psi \in \Phi_\rho \Rightarrow \phi, \psi \in \Phi_\rho$.

- $\langle a\rangle\phi \in \Phi_\rho \Rightarrow \phi \in \Phi_\rho$.
- $\phi \circ \psi \in \Phi_\rho \Rightarrow \phi, \psi \in \Phi_\rho$.
- $\phi \boldsymbol{P} \psi \in \Phi_\rho \Rightarrow \phi, \psi \in \Phi_\rho$.

10.5.1. Claim. *For any formula ρ, the subformula set of ρ is closed under subformulas.*

In ALX1, the truth condition of $\phi \boldsymbol{P} \psi$ depends on the conjunction expansion principle $cw(w, [\![\phi \wedge \neg\psi]\!]_M) \succ cw(w, [\![\psi \wedge \neg\phi]\!]_M)$. Because of this, we shall need an extended subformula set to handle the problem. We define the extended subformula set accordingly.

10.5.3. Definition. (Extended subformula set) *Let Φ be a formula set which is closed under subformulas. The extended subformula set Φ^+ is defined as the Boolean closure of Φ and select representatives for each formula up to propositional equivalence. In particular, we have $\Phi \subseteq \Phi^+$.*

It is easy to see that Φ^+ is finite, since Φ is finite. Moreover, for any $\phi, \psi \in \Phi^+$, there exist formulas χ_1 and χ_2 such that
$\chi_1 \in \Phi^+$ and $\chi_2 \in \Phi^+$ and $\vdash (\chi_1 \leftrightarrow (\phi \wedge \neg\psi))$ and $\vdash (\chi_2 \leftrightarrow (\psi \wedge \neg\phi))$.

In particular, we have $\bot \in \Phi^+$.

10.5.4. Definition. (Equivalence relation on possible worlds) *Let M be an ALX1 model $\langle W, \succ, cw, R^a, V\rangle$ and Φ be a formula set which is closed under subformulas. For any $w, w' \in W$, we define*

$w \approx w'$ with respect to M and Φ^+ iff $\forall \rho \in \Phi^+ (M, w \Vdash \rho \Leftrightarrow M, w' \Vdash \rho)$

10.5.5. Definition. (Equivalence class) *Let $\approx$ be an equivalence relation on possible worlds with respect to M and Φ^+, and w be a possible world, we define*

$[w] \stackrel{\text{def}}{=} \{w' \in W \ : \ w \approx w'\}$

10.5.6. Definition. (Filtration) *A filtration of $M = \langle W, \succ, cw, R^a, V\rangle$ through Φ^+ is any model $M^* = \langle W^*, \succ^*, cw^*, R^{a*}, V^*\rangle$ which satisfies the following conditions:*

(1) W^* is a subset of W which consists of exactly one world from each equivalence class.

(2) $R^{a*}, cw^*, \succ^*$ satisfy the following suitability conditions:

(2.1) $\forall w, w' \in W^* ((\exists u \in W)(R^a wu$ and $w' \approx u) \Rightarrow R^{a*} ww')$.

(2.2) $\forall w, w' \in W^* (R^{a*} ww' \Rightarrow (\forall \langle a\rangle\phi \in \Phi^+)(M, w' \Vdash \phi \Rightarrow M, w \Vdash \langle a\rangle\phi))$.

(2.3) $\forall w, w' \in W^*(\forall \psi \in \Phi^+)(((\exists u \in W)(w \in cw(u, [\![\psi]\!]_M)$ *and* $w' \approx u) \Rightarrow w \in cw^*(w', [\![\psi]\!]_{M^*}))$.

(2.4) $\forall w, w' \in W^*((\forall \psi \in \Phi^+)(w \in cw^*(w', [\![\psi]\!]_{M^*})$
$\Rightarrow (\forall \phi \in \Phi^+)((M, w' \Vdash \phi \wedge \psi \Rightarrow M, w \Vdash (\phi \wedge \psi) \circ \psi)$
and $(M, w' \Vdash \phi \wedge \neg\psi \Rightarrow M, w \Vdash (\phi \wedge \neg\psi) \circ \psi)$
and $(M, w' \Vdash (\neg\phi \wedge \psi) \Rightarrow M, w \Vdash (\neg\phi \wedge \psi) \circ \psi)$
and $(M, w' \Vdash (\neg\phi \wedge \neg\psi) \Rightarrow M, w \Vdash (\neg\phi \wedge \neg\psi) \circ \psi)))$.

(2.5) $\forall w, \in W^*(\forall \phi, \psi \in \Phi^+)(cw^*(w, [\![\phi \wedge \neg\psi]\!]_{M^*}) \succ^* cw^*(w, [\![\neg\phi \wedge \psi]\!]_{M^*})) \Leftrightarrow M, w \Vdash (\phi \boldsymbol{P} \psi))$.

(3) $V^*(p_i) = V(p_i)$ *for any* $p_i \in \Phi^+$.

10.5.1. Theorem. (Filtration theorem) *Let* $M = \langle W, \succ, cw, R^a, V\rangle$ *be any ALX1 model,* Φ *be any formula set which is closed under subformulas, and* $M^* = \langle W^*, \succ^*, cw^*, R^{a*}, V^*\rangle$ *be any filtration of* M *through* Φ^+*, then for any* $\chi \in \Phi^+$ *, and any* $w \in W^*(M, w \Vdash \chi \Leftrightarrow M^*, w \Vdash \chi)$.

PROOF. We prove the theorem by induction on the complexity of Φ^+.

For any χ such that $\chi \in \Phi^+$ and $\chi \in \Phi$,

Cases (1) $\chi \equiv p_i$, (2) $\chi \equiv \neg\phi$, and (3) $\chi \equiv \phi \wedge \psi$, are straightforward.

(4) $\chi \equiv \langle a\rangle\phi$. We know that $\phi \in \Phi^+$.

	$M, w \Vdash \langle a\rangle\phi$	
$\Rightarrow$	$\exists u \in W(R^a wu$ and $M, u \Vdash \phi)$	(Truth condition)
$\Rightarrow$	$\exists w' \in W^*(R^a wu$ and $w' \approx u$ and $M, u \Vdash \phi)$	(Definition of W^*)
$\Rightarrow$	$\exists w' \in W^*(R^a wu$ and $w' \approx u$ and $M, w' \Vdash \phi)$	(Definition of $\approx$)
$\Rightarrow$	$\exists w' \in W^*(R^{a*} ww'$ and $M, w' \Vdash \phi)$	(2.1)
$\Rightarrow$	$\exists w' \in W^*(R^{a*} ww'$ and $M^*, w' \Vdash \phi)$	(Induction hypothesis)
$\Rightarrow$	$M^*, w \Vdash \langle a\rangle\phi$	(Truth condition)

	$M^*, w \Vdash \langle a\rangle\phi$	
$\Leftrightarrow$	$\exists w' \in W^*(R^{a*} ww'$ and $w' \in [\![\phi]\!]_{M^*})$	(Truth condition)
$\Leftrightarrow$	$\exists w' \in W^*(R^{a*} ww'$ and $w' \in [\![\phi]\!]_M)$	(Induction hypothesis)
$\Rightarrow$	$M, w \Vdash \langle a\rangle\phi$	(2.2)

(5) $\chi \equiv \phi \circ \psi$. We know that $\phi, \psi \in \Phi^+$.

$$
\begin{array}{lll}
 & M, w \Vdash \phi \circ \psi & \\
\Rightarrow & \exists u(M, u \Vdash \phi \text{ and } w \in cw(u, [\![\psi]\!]_M)) & \text{(Truth condition)} \\
\Rightarrow & \exists w' \in W^*(M, u \Vdash \phi \text{ and } w' \approx u \text{ and } w \in cw(u, [\![\psi]\!]_M)) & \text{(Definition of } W^*) \\
\Rightarrow & \exists w' \in W^*(M, w' \Vdash \phi \text{ and } w \in cw(u, [\![\psi]\!]_M)) & \text{(Definition of } \approx) \\
\Rightarrow & \exists w' \in W^*(M, w' \Vdash \phi \text{ and } w \in cw^*(w', [\![\psi]\!]_{M^*})) & \text{(2.3)} \\
\Rightarrow & \exists w' \in W^*(M^*, w' \Vdash \phi \text{ and } w \in cw^*(w', [\![\psi]\!]_{M^*})) & \text{(Induction hypothesis)} \\
\Rightarrow & M^*, w \Vdash \phi \circ \psi & \text{(Truth condition)}
\end{array}
$$

$$
\begin{array}{lll}
 & M^*, w \Vdash \phi \circ \psi & \\
\Leftrightarrow & \exists w' \in W^*(M^*, w' \Vdash \phi \text{ and } w \in cw^*(w', [\![\psi]\!]_{M^*})) & \text{(Truth condition)} \\
\Leftrightarrow & \exists w' \in W^*(M, w' \Vdash \phi \text{ and } w \in cw^*(w', [\![\psi]\!]_{M^*})) & \text{(Induction hypothesis)}
\end{array}
$$

Case 1. $M, w' \Vdash \psi$.

$$
\begin{array}{lll}
 & M, w' \Vdash \psi \text{ and } M, w' \Vdash \phi \text{ and } w \in cw^*(w', [\![\psi]\!]_{M^*}) & \\
\Rightarrow & M, w' \Vdash (\phi \wedge \psi) \text{ and } w \in cw^*(w', [\![\psi]\!]_{M^*}) & \text{(Truth condition)} \\
\Rightarrow & M, w \Vdash (\phi \wedge \psi) \circ \psi & \text{(2.4)} \\
\Rightarrow & M, w \Vdash \phi \circ \psi & \text{(MONU)}
\end{array}
$$

Case 2. $M, w' \Vdash \neg\psi$.

$$
\begin{array}{lll}
 & M, w' \Vdash \neg\psi \text{ and } M, w' \Vdash \phi \text{ and } w \in cw^*(w', [\![\psi]\!]_{M^*})) & \\
\Rightarrow & M, w' \Vdash (\phi \wedge (\neg\psi)) \text{ and } w \in cw^*(w', [\![\psi]\!]_{M^*})) & \text{(Truth condition)} \\
\Rightarrow & M, w \Vdash (\phi \wedge (\neg\psi)) \circ \psi & \text{(2.4)} \\
\Rightarrow & M, w \Vdash \phi \circ \psi & \text{(MONU)}
\end{array}
$$

(6) $\chi \equiv \phi \boldsymbol{P} \psi$. We know that $\phi, \psi \in \Phi$. So $\phi, \psi \in \Phi^+$.

$$
\begin{array}{lll}
 & M, w \Vdash \phi \boldsymbol{P} \psi & \\
\Leftrightarrow & cw^*(w, [\![\phi \wedge \neg\psi]\!]_M) \succ^* cw^*(w, [\![\psi \wedge \neg\phi]\!]_{M^*}) & \text{(2.5)} \\
\Leftrightarrow & M^*, w \Vdash \phi \boldsymbol{P} \psi & \text{(Truth condition)}
\end{array}
$$

For any χ such that $\chi \in \Phi^+$ but $\chi \notin \Phi$, we know the following facts:

(7) $\chi \equiv \neg\phi$ and $\phi \in \Phi$. So $\phi \in \Phi^+$.

$$
\begin{array}{lll}
 & M, w \Vdash \neg\phi & \\
\Leftrightarrow & M, w \nVdash \phi & \text{(Truth condition)} \\
\Leftrightarrow & M^*, w \nVdash \phi & \text{(Induction hypothesis)} \\
\Leftrightarrow & M^*, w \Vdash \neg\phi & \text{(Truth condition)}
\end{array}
$$

(8) $\chi \equiv \phi \wedge \psi$ and $\phi, \psi \in \Phi$.

$$
\begin{array}{lll}
 & M, w\|\!\!-\phi \wedge \psi & \\
\Leftrightarrow & M, w\|\!\!-\phi \text{ and } M, w\|\!\!-\psi & \text{(Truth condition)} \\
\Leftrightarrow & M^*, w\|\!\!-\phi \text{ and } M^*, w\|\!\!-\psi & \text{(Induction hypothesis)} \\
\Leftrightarrow & M^*, w\|\!\!-\phi \wedge \psi & \text{(Truth condition)}
\end{array}
$$

Therefore, for any $\chi \in \Phi^+$, we have $M, w\|\!\!-\chi \leftrightarrow M^*, w\|\!\!-\chi$. □

10.5.1. Corollary. (Filtration corollary) *Let $M = \langle W, \succ, cw, R^a, V\rangle$ be any ALX1 model, Φ be any formula set which is closed under subformulas, and $M^* = \langle W^*, \succ^*, cw^*, R^{a*}, V^*\rangle$ be any filtration of M through Φ^+, then for any $\phi, \psi \in \Phi^+$, and $w \in W^*$,*

(a) $M, w\|\!\!-\neg\phi \Leftrightarrow M^*, w\|\!\!-\neg\phi$.
(b) $M, w\|\!\!-\phi \wedge \psi \Leftrightarrow M^*, w\|\!\!-\phi \wedge \psi$.
(c) $M, w\|\!\!-\phi \wedge \neg\psi \Leftrightarrow M^*, w\|\!\!-\phi \wedge \neg\psi$.

PROOF. Straightforward. □

10.5.2. Theorem. (Invalidity theorem) *Suppose that a formula χ is invalid in a model M, then χ is invalid in every filtration of M through Φ^+_χ.*

PROOF. Since χ is invalid in an ALX1 model $M = \langle W, \succ, cw, R^a, V\rangle$, there is some $w \in W$ such that $M, w\|\!\!-\neg\chi$. Suppose that $M^* = \langle W^*, cw^*, \succ^*, R^{a*}, V^*\rangle$ is a filtration of M through Φ^+_χ. By the definition of W^*, there is some $w^* \in W^*$ such that $w \approx w^*$ with respect to M and Φ^+_χ. Obviously, $\chi \in \Phi^+_\chi$, therefore, $M, w^*\|\!\!-\neg\chi$. By the last corollary (a), $M^*, w^*\|\!\!-\neg\chi$, and so χ is invalid in M^*. □

10.5.3. Theorem. *ALX1 has the finite model property.*

PROOF. For arbitrary χ, suppose that $\nvdash_{ALX1} \chi$, then there exists a model $M = \langle W, \succ, cw, R^a, V\rangle$ and a world $w \in W$ such that $M, w\|\!\!\not-\chi$. Let Φ_χ be the subformula set of χ. We know that Φ_χ is finite. Moreover, Φ^+_χ also is finite, by the definition of Φ^+_χ.

Now, we construct a filtration $M^* = \langle W^*, \succ^*, cw^*, R^{a*}, V^*\rangle$ of M through Φ^+_χ as follows:

(1) For W^*, we first construct the equivalence class [ ] on W as.

$[w] \stackrel{\text{def}}{=} \{w' : \forall\rho \in \Phi^+_\chi(M, w\|\!\!-\rho \Leftrightarrow M, w'\|\!\!-\rho)\}$

From each class $[w]$, we select exactly one world $w' \in [w]$ to represent this class. Now let W^* be the set of all representing worlds.

From the definition of the equivalence class [ ], we know that for any class $[w1]$, and any class $[w2]$, if $[w1] \neq [w2]$, then there exists $\rho \in \Phi^+_\chi$ such that either $M, w1 \Vdash \rho$ and $M, w2 \nVdash \rho$, or $M, w1 \nVdash \rho$ and $M, w2 \Vdash \rho$. Because Φ^+_χ is finite, there are only finitely many formulas ρ by which we can distinguish two different classes. Therefore, there are only finitely many equivalence classes, namely, at most $2^{Card(\Phi^+_\chi)}$. So W^* is finite.

(2) For V^*, we define $V^*(p_i) =_{df} V(p_i)$ if $p_i \in \Phi^+_\chi$.

(3) For R^{a*}, we define that, for any $w, w' \in W^*$,
$\langle w, w' \rangle \in R^{a*}$ iff $(\forall \langle a \rangle \phi \in \Phi^+_\chi)(M, w' \Vdash \phi \Rightarrow M, w \Vdash \langle a \rangle \phi)$.

(4) For cw^*, we define that, for any $w, w' \in W^*$, and any $\psi \in \Phi^+_\chi$,
$w \in cw^*(w', [\![\psi]\!]_{M^*})$ iff $((\forall \phi \in \Phi^+_\chi)((M, w' \Vdash \phi \wedge \psi \Rightarrow M, w \Vdash (\phi \wedge \psi) \circ \psi)$
and $(M, w' \Vdash \phi \wedge \neg\psi \Rightarrow M, w \Vdash (\phi \wedge \neg\psi) \circ \psi)$
and $(M, w' \Vdash (\neg\phi \wedge \psi) \Rightarrow M, w \Vdash (\neg\phi \wedge \psi) \circ \psi)$
and $(M, w' \Vdash (\neg\phi \wedge \neg\psi) \Rightarrow M, w \Vdash (\neg\phi \wedge \neg\psi) \circ \psi)))$.

(5) For $\succ^*$, we define that, for any $w \in W^*$, and any $\phi, \psi \in \Phi^+_\chi$,
$cw^*(w, [\![\phi \wedge \neg\psi]\!]_{M^*}) \succ^* cw^*(w, [\![\neg\phi \wedge \psi]\!]_{M^*})$ iff $M, w \Vdash (\phi \boldsymbol{P} \psi)$.

Now, we have to show that M^* satisfies the conditions of a filtration of M through Φ^+_χ. From the construction of W^*, we know that W^* is a subset of W. Moreover, W^* consists of exactly one world from each equivalence class with respect to M and Φ^+_χ. Therefore, the condition for W^* is satisfied.

From the above definition of $V^*, R^{a*}, cw^*, \succ^*$, the suitability conditions (2.2), (2.4), (2.5), and the condition for V^* are obviously satisfied.

To show that (2.1) is satisfied, we have to show that $\forall w, w' \in W^*((\exists u \in W)(w' \approx u$ and $R^a wu) \Rightarrow R^{a*} ww')$.

Suppose that $\exists u \in W(w' \approx u$ and $R^a wu)$, and for any $\langle a \rangle \phi \in \Phi^+_\chi$.

$$\begin{array}{lll} & \langle a \rangle \phi \in \Phi^+_\chi \text{ and } M, w' \Vdash \phi \text{ and } R^a wu & \\ \Rightarrow & \langle a \rangle \phi \in \Phi^+_\chi \text{ and } M, u \Vdash \phi \text{ and } R^a wu & \text{(Definition of } \approx\text{)} \\ \Rightarrow & \langle a \rangle \phi \in \Phi^+_\chi \text{ and } M, w \Vdash \langle a \rangle \phi & \text{(Truth condition)} \end{array}$$

Therefore, according to the definition of R^{a*}, we have $R^{a*} ww'$, so (2.1) holds.

To show that (2.3) is satisfied, we have to show that $(\forall w, w' \in W^*)(\forall \psi \in \Phi^+_\chi)((\exists u \in W)(w' \approx u$ and $w \in cw(u, [\![\psi]\!]_M)) \Rightarrow w \in cw^*(w', [\![\psi]\!]_{M^*}))$.

For any $w, w' \in W^*$, and any $\psi \in \Phi^+_\chi$, suppose that $(\exists u \in W)((w' \approx u$ and $w \in cw(u, [\![\psi]\!]_M))$, and for any $\phi \in \Phi^+_\chi$,

Assume that $M, w' \Vdash (\phi \wedge \psi)$, then

$$\begin{array}{lll} & M, w' \Vdash (\phi \wedge \psi) \text{ and } w' \approx u \text{ and } w \in cw(u, [\![\psi]\!]_M) & \\ \Rightarrow & M, u \Vdash (\phi \wedge \psi) \text{ and } w \in cw(u, [\![\psi]\!]_M) & \text{(Definition of } \approx) \\ \Rightarrow & M, w \Vdash (\phi \wedge \psi) \circ \psi & \text{(Truth condition)} \end{array}$$

Other cases $M, w' \Vdash (\phi \wedge \neg\psi)$, $(\neg\phi \wedge \psi)$, $(\neg\phi \wedge \neg\psi)$ are similar. Therefore, according to the definition of cw^* above, we have $w \in cw^*(w', [\![\psi]\!]_{M^*})$. So (2.3) holds.

We know now that M^* is indeed a filtration of M through Φ^+_χ. Moreover, we know that M^* is a finite model. By the above theorem, we know that there exists a $w \in W^*$ such that $M^*, w \nVdash \chi$. Therefore, the condition (1) of the finite model property is satisfied.

In order to show that the condition (2) for the finite model property is also satisfied, we have to show that M^* is an ALX1 model. That is to say, we have to show that cw^* satisfies (CS1), (CS2), and (CSC), and $\succ^*$ satisfies the normality and the transitivity.

For any $w, w' \in W^*$, and any $\psi \in \Phi^+_\chi$,
(CS1) $w \in cw^*(w', [\![\psi]\!]_{M^*}) \Rightarrow w \in [\![\psi]\!]_{M^*}$.

$$\begin{array}{lll} & w \in cw^*(w', [\![\psi]\!]_{M^*}) & \\ \Leftrightarrow & (\forall \phi \in \Phi^+_\chi)((M, w' \Vdash (\phi \wedge \psi) \Rightarrow M, w \Vdash (\phi \wedge \psi) \circ \psi) & \\ & \quad \text{and } (M, w' \Vdash (\phi \wedge \neg\psi) \Rightarrow M, w \Vdash (\phi \wedge \neg\psi) \circ \psi) & \\ & \quad \text{and } (M, w' \Vdash (\neg\phi \wedge \psi) \Rightarrow M, w \Vdash (\neg\phi \wedge \psi) \circ \psi) & \\ & \quad \text{and } (M, w' \Vdash (\neg\phi \wedge \neg\psi) \Rightarrow M, w \Vdash (\neg\phi \wedge \neg\psi) \circ \psi)) & \text{(Definition of } cw^*) \end{array}$$

Case 1 $M, w' \Vdash (\phi \wedge \psi)$.

$$\begin{array}{lll} & M, w' \Vdash (\phi \wedge \psi) & \\ \Rightarrow & M, w \Vdash (\phi \wedge \psi) \circ \psi & \text{(Definition of } cw^*) \\ \Rightarrow & M, w \Vdash \psi & (M \text{ is an ALX model, and (U1)}) \\ \Rightarrow & M^*, w \Vdash \psi & \text{(Filtration theorem)} \\ \Rightarrow & w \in [\![\psi]\!]_{M^*} & \text{(Definition of } [\![\ ]\!]_{M^*}) \end{array}$$

The other cases $(\phi \wedge \neg\psi, \neg\phi \wedge \psi, \neg\phi \wedge \neg\psi)$ are proved similarly. Therefore, (CS1) is satisfied.

(CS2) $w \in [\![\psi]\!]_{M^*} \Rightarrow cw^*(w, [\![\psi]\!]_{M^*}) = \{w\}$.

we must show that:
(a) $w \in [\![\psi]\!]_{M^*} \Rightarrow w \in cw^*(w, [\![\psi]\!]_{M^*})$.
(b) $w \in [\![\psi]\!]_{M^*}$ and $w' \in cw^*(w, [\![\psi]\!]_{M^*}) \Rightarrow w \equiv w'$.

Where $w \equiv w'$ means that w and w' represent the same equivalence class with respect to M^* and Φ^+_χ, namely,

$\forall \rho \in \Phi^+_\chi (M^*, w \Vdash \rho \Leftrightarrow M^*, w \Vdash \rho)$

For (a), we will show that $w \in [\![\psi]\!]_{M^*}$ and $w \notin cw^*(w, [\![\psi]\!]_{M^*}) \Rightarrow$ **False**

$$
\begin{array}{lll}
 & w \in [\![\psi]\!]_{M^*} \text{ and } w \notin cw^*(w, [\![\psi]\!]_{M^*}) & \\
\Rightarrow & w \in [\![\psi]\!]_{M} \text{ and } w \notin cw^*(w, [\![\psi]\!]_{M^*}) & \text{(Filtration theorem)} \\
\Rightarrow & w \in [\![\psi]\!]_{M} \text{ and } (\exists \phi \in \Phi^+_\chi)((M, w \Vdash (\phi \wedge \psi) \text{ and} & \\
 & M, w \nVdash (\phi \wedge \psi) \circ \psi) & \\
 & \quad \text{or } (M, w \Vdash (\phi \wedge \neg\psi) \text{ and } M, w \nVdash (\phi \wedge \neg\psi) \circ \psi) & \\
 & \quad \text{or } (M, w \Vdash (\neg\phi \wedge \psi) \text{ and } M, w \nVdash (\neg\phi \wedge \psi) \circ \psi) & \\
 & \quad \text{or } (M, w \Vdash (\neg\phi \wedge \neg\psi) \text{ and } M, w \nVdash (\neg\phi \wedge \neg\psi) \circ \psi)) & \text{(Definition of } cw^*) \\
\Rightarrow & w \in [\![\psi]\!]_{M} \text{ and } (\exists \phi \in \Phi^+_\chi)((M, w \Vdash (\phi \wedge \psi) \text{ and} & \\
 & M, w \nVdash (\phi \wedge \psi) \circ \psi) & \\
 & \quad \text{or } (M, w \Vdash (\neg\psi) \text{ and } M, w \nVdash (\phi \wedge \neg\psi) \circ \psi) & \\
 & \quad \text{or } (M, w \Vdash \neg\phi \wedge \psi \text{ and } M, w \nVdash (\neg\phi \wedge \psi) \circ \psi) & \\
 & \quad \text{or } (M, w \Vdash (\neg\psi) \text{ and } M, w \nVdash (\neg\phi \wedge \neg\psi) \circ \psi)) & \text{(MONU)}
\end{array}
$$

$$
\begin{array}{lll}
\Rightarrow & w \in [\![\psi]\!]_{M} \text{ and } (\exists \phi \in \Phi^+_\chi)((M, w \Vdash (\phi \wedge \psi) \text{ and} & \\
 & M, w \nVdash (\phi \wedge \psi) \circ \psi) & \\
 & \quad \text{or } (M, w \Vdash (\neg\phi \wedge \psi) \text{ and} & \\
 & M, w \nVdash (\neg\phi \wedge \psi) \circ \psi)) & \text{(Meta reasoning)} \\
\Rightarrow & \exists \phi \in \Phi^+_\chi ((M, w \Vdash ((\phi \wedge \psi) \wedge \psi) \text{ and } M, w \nVdash (\phi \wedge \psi) \circ \psi) & \\
 & \text{or } (M, w \Vdash ((\neg\phi \wedge \psi) \wedge \psi) \text{ and } M, w \nVdash (\neg\phi \wedge \psi) \circ \psi)) & (M, w \Vdash \psi) \\
\Rightarrow & \exists \phi \in \Phi^+_\chi ((M, w \Vdash ((\phi \wedge \psi) \circ \psi) \text{ and } M, w \nVdash (\phi \wedge \psi) \circ \psi) & \\
 & \quad \text{or } (M, w \Vdash ((\neg\phi \wedge \psi) \circ \psi) \text{ and } M, w \nVdash (\neg\phi \wedge \psi) \circ \psi)) & \text{(U2)} \\
\Rightarrow & \textbf{False} &
\end{array}
$$

For (b), suppose that $w \in [\![\psi]\!]_{M^*}$ and $w' \in cw^*(w, [\![\psi]\!]_{M^*})$, for any $\rho \in \Phi^+_\chi$, we have to show that

$M^*, w \Vdash \rho \Leftrightarrow M^*, w' \Vdash \rho$.

($\Rightarrow$) We show that $M^*, w \Vdash \rho$ and $M^*, w' \nVdash \rho \Rightarrow$ **False**.

$$
\begin{array}{lll}
 & M^*, w \| \!\!- \rho \text{ and } M^*, w' \| \!\!\not\!\!- \rho & \\
\Rightarrow & M, w \| \!\!- \rho \text{ and } M^*, w' \| \!\!\not\!\!- \rho & \text{(Filtration theorem)} \\
\Rightarrow & M, w \| \!\!- \rho \text{ and } M^*, w' \| \!\!- \neg\rho & \text{(Truth condition)} \\
\Rightarrow & M, w \| \!\!- \rho \text{ and } M, w' \| \!\!- \neg\rho & \text{(Filtration corollary (a))} \\
\Rightarrow & M, w \| \!\!- \rho \text{ and } M, w' \| \!\!- \neg((\rho \wedge \psi) \circ \psi) & (M \text{ is an ALX1 model}) \\
\Rightarrow & M, w \| \!\!- (\rho \wedge \psi) \text{ and } M, w' \| \!\!- \neg((\rho \wedge \psi) \circ \psi) & (w \in [\![\psi]\!]_{M^*}) \\
\Rightarrow & M, w' \| \!\!- (\rho \wedge \psi) \circ \psi \text{ and } M, w' \| \!\!- \neg((\rho \wedge \psi) \circ \psi) & \text{(Definition of } cw^*) \\
\Rightarrow & \textbf{False} &
\end{array}
$$

($\Leftarrow$) We show that $M^*, w' \| \!\!- \rho$ and $M^*, w \| \!\!\not\!\!- \rho \Rightarrow$ **False**.

$$
\begin{array}{lll}
 & M^*, w' \| \!\!- \rho \text{ and } M^*, w \| \!\!\not\!\!- \rho & \\
\Rightarrow & M, w' \| \!\!- \rho \text{ and } M^*, w \| \!\!\not\!\!- \rho & \text{(Filtration theorem)} \\
\Rightarrow & M, w' \| \!\!- \rho \text{ and } M^*, w \| \!\!- \neg\rho & \text{(Truth condition)} \\
\Rightarrow & M, w' \| \!\!- \rho \text{ and } M, w \| \!\!- \neg\rho & \text{(Filtration corollary (a))} \\
\Rightarrow & M, w' \| \!\!- \rho \text{ and } M^*, w \| \!\!- \psi \text{ and } M, w \| \!\!- \neg\rho & (w \in [\![\psi]\!]_{M^*}) \\
\Rightarrow & M, w' \| \!\!- \rho \text{ and } M, w \| \!\!- \psi \text{ and } M, w \| \!\!- \neg\rho & \text{(Filtration theorem)} \\
\Rightarrow & M, w' \| \!\!- \rho \text{ and } M, w \| \!\!- (\neg\rho \wedge \psi) & \text{(Truth condition)} \\
\Rightarrow & M, w' \| \!\!- \rho \text{ and } M, w' \| \!\!- (\neg\rho \wedge \psi) \circ \psi & \text{(Definition of } cw^*) \\
\Rightarrow & M, w' \| \!\!- \neg(\neg\rho) \text{ and } M, w' \| \!\!- (\neg\rho \wedge \psi) \circ \psi & \text{(Meta reasoning)} \\
\Rightarrow & M, w' \| \!\!- \neg(((\neg\rho) \wedge \psi) \circ \psi)) \text{ and } M, w' \| \!\!- (\neg\rho \wedge \psi) \circ \psi & (M \text{ is an ALX1 model}) \\
\Rightarrow & \textbf{False} &
\end{array}
$$

(CSC) $j \in cw^*(w, [\![\phi]\!]_{M^*}) \cap [\![\psi]\!]_{M^*} \Rightarrow j \in cw^*(w, [\![\phi \wedge \psi]\!]_{M^*})$.

$$
\begin{array}{lll}
 & j \in cw^*(w, [\![\phi]\!]_{M^*}) \cap [\![\psi]\!]_{M^*} & \\
\Leftrightarrow & (\forall \rho \in \Phi^+_\chi)((M, w \| \!\!- (\rho \wedge \phi) \Rightarrow M, j \| \!\!- (\rho \wedge \phi) \circ \phi) & \\
 & \quad \text{and } (M, w \| \!\!- (\rho \wedge \neg\phi) \Rightarrow M, j \| \!\!- (\rho \wedge \neg\phi) \circ \phi) & \\
 & \quad \text{and } (M, w \| \!\!- (\neg\rho \wedge \phi) \Rightarrow M, j \| \!\!- (\neg\rho \wedge \phi) \circ \phi) & \\
 & \quad \text{and } (M, w \| \!\!- (\neg\rho \wedge \neg\phi) \Rightarrow M, j \| \!\!- (\neg\rho \wedge \neg\phi) \circ \phi)) & \\
 & \text{and } M^*, j \| \!\!- \psi & \text{(Definition of } cw^*) \\
\Rightarrow & M, j \| \!\!- \psi & \text{(Filtration lemma)}
\end{array}
$$

For any $\rho \in \Phi^+_\chi$,

Case 1 $M, w \| \!\!- (\rho \wedge (\phi \wedge \psi))$.

$$
\begin{array}{lll}
 & M, w \| \!\!- (\rho \wedge (\phi \wedge \psi) & \\
\Rightarrow & M, w \| \!\!- (\rho \wedge \phi) & \text{(Meta reasoning)} \\
\Rightarrow & M, j \| \!\!- (\rho \wedge \phi) \circ \phi & (j \in cw^*(w, [\![\phi]\!]_{M^*})) \\
\Rightarrow & M, j \| \!\!- (\rho \wedge (\phi \wedge \psi)) \circ (\phi \wedge \psi) & (M, j \| \!\!- \psi \text{ and (U8}^\circ)\text{)}
\end{array}
$$

Case 2 $M, w\|\!\!-(\rho \wedge \neg(\phi \wedge \psi)) \Rightarrow M, w\|\!\!-(\rho \wedge \neg\phi) \vee (\rho \wedge \neg\psi)$.

Case 2.1

	$M, w\|\!\!-(\rho \wedge \neg\phi)$	
$\Rightarrow$	$M, j\|\!\!-(\rho \wedge \neg\phi) \circ \phi$	$(j \in cw^*(w, [\![\phi]\!]_{M^*}))$
$\Rightarrow$	$M, j\|\!\!-(\rho \wedge \neg\phi) \circ (\phi \wedge \psi)$	$(M, j\|\!\!-\psi$ and (U6))
$\Rightarrow$	$M, j\|\!\!-(\rho \wedge \neg\phi \vee \rho \wedge \neg\psi) \circ (\phi \wedge \psi)$	(MONU)
$\Rightarrow$	$M, j\|\!\!-(\rho \wedge \neg(\phi \wedge \psi)) \circ (\phi \wedge \psi)$	(Meta reasoning)

Case 2.2 $M, w\|\!\!-\rho \wedge \neg\psi$

Case 2.2.1

	$M, w\|\!\!-\phi$	
$\Rightarrow$	$M^*, w\|\!\!-\phi$	(Filtration lemma)
$\Rightarrow$	$\{w\} = cw^*(w, [\![\phi]\!]_{M^*})$	(CS2)
$\Rightarrow$	$w = j$	$(j \in cw^*(w, [\![\phi]\!]_{M^*}))$
$\Rightarrow$	$M, j\|\!\!-\neg\psi$	$(M, w\|\!\!-\neg\psi)$
$\Rightarrow$	**False**	$(M, j\|\!\!-\psi)$

Case 2.2.2

	$M, w\|\!\!-\neg\phi$	
$\Rightarrow$	$M, w\|\!\!-(\rho \wedge \neg\phi)$	(Meta reasoning)
$\Rightarrow$	$M, j\|\!\!-(\rho \wedge \neg\phi) \circ \phi$	$(j \in cw^*(w, [\![\phi]\!]_{M^*}))$
$\Rightarrow$	$M, j\|\!\!-(\rho \wedge \neg\phi) \circ (\phi \wedge \psi)$	(U6)
$\Rightarrow$	$M, j\|\!\!-(\rho \wedge \neg\phi \vee \rho \wedge \neg\psi) \circ (\phi \wedge \psi)$	(MONU)
$\Rightarrow$	$M, j\|\!\!-(\rho \wedge \neg(\phi \wedge \psi)) \circ (\phi \wedge \psi)$	(Meta reasoning)

Case 3

	$M, w\|\!\!-\neg\rho \wedge (\phi \wedge \psi)$	
$\Rightarrow$	$M, w\|\!\!-\neg\rho \wedge \phi$	(Meta reasoning)
$\Rightarrow$	$M, j\|\!\!-(\neg\rho \wedge \phi) \circ \phi$	$(j \in cw^*(w, [\![\phi]\!]_{M^*}))$
$\Rightarrow$	$M, j\|\!\!-(\neg\rho \wedge (\phi \wedge \psi) \circ \phi$	$(M, j\|\!\!-\psi$, and (U8°))
$\Rightarrow$	$M, j\|\!\!-(\neg\rho \wedge (\phi \wedge \psi)) \circ (\phi \wedge \psi)$	(U4)

Case 4 $M, w\|\!\!-\neg\rho \wedge \neg(\phi \wedge \psi) \Rightarrow M, w\|\!\!-(\neg\rho \wedge \neg\phi) \vee (\rho \wedge \neg\psi)$.

Case 4.1

	$M, w\|\!\!-(\neg\rho \wedge \neg\phi)$	
$\Rightarrow$	$M, j\|\!\!-(\neg\rho \wedge \neg\phi) \circ \phi$	$(j \in cw^*(w, [\![\phi]\!]_{M^*}))$
$\Rightarrow$	$M, j\|\!\!-(\neg\rho \wedge \neg\phi) \circ (\phi \wedge \psi)$	$(M, j\|\!\!-\psi$ and (U6))
$\Rightarrow$	$M, j\|\!\!-(\neg\rho \wedge \neg\phi \vee \rho \wedge \neg\psi) \circ (\phi \wedge \psi)$	(MONU)
$\Rightarrow$	$M, j\|\!\!-(\neg\rho \wedge \neg(\phi \wedge \psi)) \circ (\phi \wedge \psi)$	(Meta reasoning)

Case 4.2 $M, w \Vdash \neg\rho \wedge \neg\psi$

Case 4.2.1

	$M, w \Vdash \phi$	
$\Rightarrow$	$M^*, w \Vdash \phi$	(Filtration lemma)
$\Rightarrow$	$\{w\} = cw^*(w, [\![\phi]\!]_{M^*})$	(CS2)
$\Rightarrow$	$w = j$	$(j \in cw^*(w, [\![\phi]\!]_{M^*}))$
$\Rightarrow$	$M, j \Vdash \neg\psi$	$(M, w \Vdash \neg\psi)$
$\Rightarrow$	**False**	$(M, j \Vdash \psi)$

Case 4.2.2

	$M, w \Vdash \neg\phi$	
$\Rightarrow$	$M, w \Vdash (\neg\rho \wedge \neg\phi)$	(Meta reasoning)
$\Rightarrow$	$M, j \Vdash (\neg\rho \wedge \neg\phi) \circ \phi$	$(j \in cw^*(w, [\![\phi]\!]_{M^*}))$
$\Rightarrow$	$M, j \Vdash (\neg\rho \wedge \neg\phi) \circ (\phi \wedge \psi)$	(U6)
$\Rightarrow$	$M, j \Vdash (\neg\rho \wedge \neg\phi \vee \rho \wedge \neg\psi) \circ (\phi \wedge \psi)$	(MONU)
$\Rightarrow$	$M, j \Vdash (\neg\rho \wedge \neg(\phi \wedge \psi)) \circ (\phi \wedge \psi)$	(Meta reasoning)

Therefore, by the results of the cases 1-4 and the definition of cw^*, we have that $j \in cw^*(w, [\![\phi \wedge \psi]\!]_{M^*})$. So (CSC) is satisfied.

(NORM) $(\emptyset \not\succ^* X)$.
We must show that $\emptyset \succ^* X \Rightarrow$ **False**.

	$\emptyset \succ^* X$	
$\Rightarrow$	$(\exists\phi, \psi \in \Phi^+_\chi)(M, w \Vdash \phi \boldsymbol{P} \psi$ and $cw^*(w, [\![\phi \wedge \neg\psi]\!]_{M^*}) = \emptyset$ and $cw^*(w, [\![\psi \wedge \neg\phi]\!]_{M^*}) = X$ and $cw^*(w, [\![\phi \wedge \neg\psi]\!]_{M^*}) \succ^* cw^*(w, [\![\psi \wedge \neg\phi]\!]_{M^*}))$	(Definition of $\succ$)
$\Rightarrow$	$cw^*(w, [\![\bot]\!]_{M^*}) \succ^* cw^*(w, [\![\psi \wedge \neg\phi]\!]_{M^*})$	$(cw^*(w, [\![\bot]\!]_{M^*}) = \emptyset)$
$\Rightarrow$	$cw^*(w, [\![\bot]\!]_{M^*}) \succ^* cw^*(w, [\![\psi \wedge \neg\phi]\!]_{M^*})$ and $\bot \in \Phi^+_\chi$ and $\exists\rho \in \Phi^+_\chi([\![\rho]\!]_{M^*} = [\![\psi \wedge \neg\phi]\!]_{M^*})$	(Definition of Φ^+_χ)
$\Rightarrow$	$cw^*(w, [\![\bot \wedge \neg\rho]\!]_{M^*}) \succ^* cw^*(w, [\![\rho \wedge \neg\bot]\!]_{M^*})$	(Meta reasoning)
$\Rightarrow$	$M, w \Vdash \bot \boldsymbol{P} \rho$	(Definition of cw^*)
$\Rightarrow$	**False**	(N)

The proof for the second part of (NORM), $(X \not\succ^* \emptyset)$, is similar.

(TRAN) $cw^*(w, X \cap \overline{Y}) \succ^* cw^*(w, Y \cap \overline{X})$ and $cw^*(w, Y \cap \overline{Z}) \succ^* cw^*(w, Z \cap \overline{Y}) \Rightarrow$ $cw^*(w, X \cap \overline{Z}) \succ^* cw^*(w, Z \cap \overline{X})$.

$$
\begin{array}{lll}
& cw^*(w, X \cap \overline{Y}) \succ^* cw^*(w, Y \cap \overline{X}) \text{ and} & \\
& cw^*(w, Y \cap \overline{Z}) \succ^* cw^*(w, Z \cap \overline{Y}) & \\
\Rightarrow & \exists\phi\exists\psi\exists\rho(X = [\![\phi]\!]_{M*} \text{ and } Y = [\![\psi]\!]_{M*} \text{ and } Z = [\![\rho]\!]_{M*} & \\
& \text{and } M, w\|\!-(\phi \boldsymbol{P} \psi) \text{ and } M, w\|\!-(\psi \boldsymbol{P} \rho) & \\
& \text{and } (\phi, \psi, \rho \in \Phi^+_\chi)) & \text{(Definition of } cw^*) \\
\Rightarrow & M, w\|\!-\phi \boldsymbol{P} \rho \text{ and } (\phi, \rho \in \Phi^+_\chi) & \text{(TR)} \\
\Rightarrow & cw^*(w, [\![\phi \wedge \neg\rho]\!]_{M^*}) \succ^* cw^*(w, [\![\rho \wedge \neg\phi]\!]_{M^*}) & \text{(Definition of } \succ^*) \\
\Rightarrow & cw^*(w, X \cap \overline{Z}) \succ^* cw^*(w, Z \cap \overline{X}) & \text{(Definitions of } X, Y, Z)
\end{array}
$$

As a consequence, M^* is an ALX1 model. Because of the soundness of ALX1 logic, we know that for any ρ, $\vdash_{ALX1} \rho \Rightarrow \forall w(M^*, w\|\!-\rho)$. That means that $ALX1$ logic also satisfies the condition (2) of the finite model property. So $ALX1$ has the finite model property. □

10.6 Discussion

ALX1 provides the skeleton of a preference-driven action logic. With the completeness-results for ALX1, we have presented the first complete logic for normal preference relations, i.e., two-place relations expressing a comparative statement such as $\phi \boldsymbol{P} \psi$.

10.6.1 Goodness, Badness and Indifference

As von Wright already remarked that the notion of preference provides a basis for defining "goodness", "badness", and "indifference" [von Wright 1963].

10.6.1. Definition. (Goodness, badness, indifference) *Let Goodϕ stand for the fact that situation ϕ is perceived as good by an agent, Badϕ for the fact that situation ϕ is perceived as bad, and Indϕ that the agent is indifferent with respect to ϕ. Define:*

$$
\begin{array}{lll}
(Gdf) & Good\phi & \stackrel{\text{def}}{\Longleftrightarrow} \quad \phi \boldsymbol{P} \neg\phi. \\
(Bdf) & Bad\phi & \stackrel{\text{def}}{\Longleftrightarrow} \quad \neg\phi \boldsymbol{P} \phi. \\
(Idf) & Ind\phi & \stackrel{\text{def}}{\Longleftrightarrow} \quad \neg(\phi \boldsymbol{P} \neg\phi) \wedge \neg(\neg\phi \boldsymbol{P} \phi).
\end{array}
$$

10.6.1. Proposition. (More properties of goodness, badness and indifference)
(a) $\phi \boldsymbol{P} \psi \wedge Good\psi \rightarrow Good\phi$.
(b) $\phi \boldsymbol{P} \psi \wedge Bad\phi \rightarrow Bad\psi$.
(c) $Good\phi \leftrightarrow Bad\neg\phi$.
(d) $Good\phi \rightarrow \neg Bad\phi$.
(e) $Bad\phi \rightarrow \neg Bad\neg\phi$.
(f) $Ind\phi \leftrightarrow \neg Good\phi \wedge \neg Bad\phi$.

PROOF. (a)

$$\begin{array}{lll} & \vdash \phi \boldsymbol{P} \psi \wedge \text{Good}\psi & \\ \Rightarrow & \vdash \phi \boldsymbol{P} \psi \wedge \psi \boldsymbol{P} \neg \psi & \\ \Rightarrow & \vdash \phi \boldsymbol{P} \psi \wedge \phi \boldsymbol{P} \neg \psi & \text{(TR)} \\ \Rightarrow & \vdash \phi \boldsymbol{P} \psi \wedge \psi \boldsymbol{P} \neg \phi & \text{(CP)} \\ \Rightarrow & \vdash \phi \boldsymbol{P} \neg \phi & \text{(TR)} \\ \Rightarrow & \vdash \text{Good}\phi & \end{array}$$

The proof for (b) is similar to the proof of (a). (c)-(f) are straightforward from the definitions. □

10.6.2 Preferences

As opposed to other action logics, rational choice in ALX1 is driven by preferences. Preference logic was introduced by [Halldén 1957] and codified by [von Wright 1963, von Wright 1972], who introduced the conjunction expansion principle. Preference logic met resistance from its very beginning as numerous counterexamples against its principles, notably conjunction expansion and contraposition, were suggested [Chisholm&Sosa 1966a, Mullen 1979]. It seems reasonable to assume that these counterexamples have something to do with preference logic's failure to catch on. In the last chapter, we've dealt with the relevant counterexamples and argued that all of them are overstretching the expressive power of propositional logic. We can be more formal with respect to conjunction expansion here. The counterexamples against conjunction expansion come out false in ALX1, because

$$(\phi \wedge \psi)\boldsymbol{P}\phi \leftrightarrow \bot \boldsymbol{P}(\phi \wedge \neg\psi) \leftrightarrow \bot$$

is a theorem of ALX1 (thanks to (N), the normality axiom).

The most convincing counterexample against conjunction expansion **(CEP)** has been given by [Chisholm&Sosa 1966a]: Assume that it is better that Smith *and* his wife are happy ($p \wedge q$), than that Smith is happy on his own: $(p \wedge q)\boldsymbol{P}p$. Conjunction expansion yields $(p \wedge q \wedge \neg p)\boldsymbol{P}\neg(p \wedge q) \wedge p$, and hence the preference for a contradictory state of affairs.

In $ALX1$, this statement always comes out false. ALX1 forces the user to make the implication explicit that the happyness of the Smith couple, when compared to the happyness of Smith alone, means that if Smith's is happy alone, his wife is not happy: $(p \wedge q)\boldsymbol{P}(p \wedge \neg q)$. This statement entails no preference for a contradictory state of affairs; it is equivalent to its conjunction expansion. Other counterexamples to the conjunction expansion principle use the same trick [Chisholm&Sosa 1966a, Hansson 1968], and are hence blocked in the same way in $ALX1$.

ALX1 has a situational semantics for preference relations: the agent is supposed to have a preference of ϕ above ψ iff she would prefer ϕ-and-not-ψ to ψ-and-not-ϕ under conditions as similar as possible to her actual situation. Obvi-

ously, situational preferences can be unstable; the agent may have a specific preference in one situation and an opposite preference in another situation. Unstable preferences play an important role in many applications of bounded rationality [Carley 1986, March 1976, March&Olsen 1986, Padgett 1980], yet one might want to impose stable preferences for theoretical reasons (e.g., when one is using the logic to model economic theories where stability of preferences is often assumed [French 1988]). A stable preference relation would be one that does not change from world to world. We have discussed stable preferences in the section 9.2.3; stability of the preference relation does obtain, for example, if a preference depends only on a finite (possible empty) set of conditions that can be expressed as propositions. Stable preferences can be characterized with the following axiom:

$$(UOP) \qquad (\phi \boldsymbol{P} \psi) \circ \chi \rightarrow (\phi \boldsymbol{P} \psi).$$

10.6.3 Minimal Change and Actions

We have used the notion of minimal change to reflect the context dependency of preference statements, but minimal change may serve other purposes as well. Stalnaker introduced minimal change to modal logic to capture the semantics of the intensional conditional, e.g., counterfactual conditionals that reflect causality [Stalnaker 1968]. Since actions entail causal effects, one might also want to employ minimal change in the semantics of the action operator [Ginsberg&Smith 1987, Winslett 1988, Jackson 1989]. This could address some nastier problems of action logics, in particular the *qualification- frame-* and *ramification* problem. Actions may require a specific context for execution (qualification) that the action description must take into account. Linking actions to minimal change can provide an implicit qualification of the context of a specific action through the accessibility relation for that action; using minimal change, this context is given by the actual state (that either will or will not permit action a to be executed); no additional specification of the context is required, once the accessibility relation for action a is given. The *frame of change* is given by those conditions that do not have to change as a function of a's execution. And the ramifications of an action are "automatically" captured by identifying the set of its weakest postconditions.

ALX1 did not put strong constraints on the closest world function. Stronger constraints might be desirable, perhaps even a full-fledged definition. We have refrained from defining the closest world function in this chapter for two reasons. *First*, we wanted to provide a "logicians logic," i.e., a logic whose formal semantics is more than a faithful mirror of its syntax. As a consequence, we have used the standard semantic setup without strong restrictions on the definition of models, and have not given a circumscription of the properties of possible worlds. A definition of the closest world function would require such a circumscription. *Second*, we are not sure about the exact meaning of the notion of "closest worlds", despite various attempts in the literature to provide such a definition [Ginsberg&Smith 1987, Hansson 1989,

Jackson 1989, Winslett 1988]. There are two main problems. First, the definitions do not restrict the set of closest worlds as much as one's intuition seems to require (so the set contains more worlds than it should), and second, the definitions do not clearly distinguish between epistemically closest worlds and causally closests worlds. Such a distinction is required, however, because the closest accessible world might very well be further away than the closest imaginable world, and this difference is, indeed, important for an action logic.

We can illustrate this point by looking at minimal change as a consequence of an action. After all, the standard interpretation of actions is in terms of causality, hence in terms of minimal change. Unfortunately, there are several ways to conceptualize minimal change with respect to actions. We use $\langle a \rangle^{\#}\phi$ to denote the set of worlds where, by doing action a, the agent can achieve a minimally different ϕ-situation.

Minimal Change Actions Focusing on Accessible Worlds

One way to conceptualize such a change would be in terms of the closest world accessible via action a. Call this kind of "minimal change action" $\langle a \rangle^{\#1}$. The corresponding truth condition is:

$$[\![\langle a \rangle^{\#1}\phi]\!]_M = \{w : \exists w' \in W(w' \in cw(w, \{x : R^a wx\}) \text{ and } w' \in [\![\phi]\!]_M)\}.$$

According to this truth condition, $\langle a \rangle^{\#1}\phi$ first looks at the closest worlds accessible via action a and from this set pick the ϕ-worlds.

Consider, as an example, that a denotes the action of "slamming the door", and assume that *slamming* the door will cause the picture to fall off the wall (as opposed to, say, "closing the door" that will leave the picture unharmed). Assume ϕ stands for the fact that the door is shut. $\langle a \rangle^{\#1}\phi$ now looks at a world where the door is shut and the picture fell off the wall.
We can define:

$$R^{a\#}ww' \stackrel{\text{def}}{=} w' \in cw(w, \{x : R^a wx\}).$$

Then, the truth condition is:

$$[\![\langle a \rangle^{\#1}\phi]\!]_M = \{w : \exists w' \in W(R^{a\#}ww' \text{ and } w' \in [\![\phi]\!]_M)\}.$$

We can use a picture to convey the intuitive meaning of $\langle a \rangle^{\#1}\phi$.
We define that $[a]^{\#1}\phi$ as $\neg\langle a \rangle^{\#1}\neg\phi$.
We can show that the following axioms and inference rules for $\langle a \rangle^{\#1}$ are valid on the

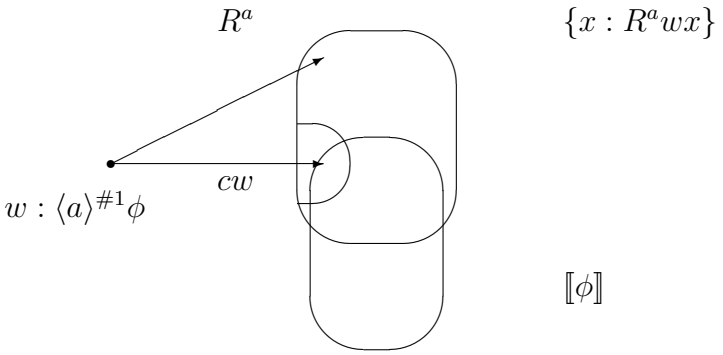


Figure 10.1: Minimal change actions focusing on accessible worlds

class of ALX1 models.

$(\boldsymbol{A1\#1})$:	$\langle a\rangle^{\#1}\bot$	$\leftrightarrow$	$\bot$.
$(\boldsymbol{A2\#1})$:	$\langle a\rangle^{\#1}(\phi \vee \psi)$	$\leftrightarrow$	$\langle a\rangle^{\#1}\phi \ \vee \langle a\rangle^{\#1}\psi$.
$(\boldsymbol{A3\#1})$:	$\langle a\rangle^{\#1}(\phi \wedge \psi)$	$\rightarrow$	$\langle a\rangle^{\#1}\phi \ \wedge \langle a\rangle^{\#1}\psi$.
$(\boldsymbol{A\#1})$:	$\langle a\rangle^{\#1}\phi$	$\rightarrow$	$\langle a\rangle\phi$.
$(\boldsymbol{NECA\#1})$:	$\vdash \phi$	$\Rightarrow$	$\vdash [a]^{\#1}\phi$.
$(\boldsymbol{MONA\#1})$:	$\vdash \langle a\rangle^{\#1}\phi$ & $\vdash \phi \rightarrow \psi$	$\Rightarrow$	$\vdash \langle a\rangle^{\#1}\psi$.
$(\boldsymbol{SUBA\#1})$:	$\vdash (\phi \leftrightarrow \phi')$	$\Rightarrow$	$\vdash (\langle a\rangle^{\#1}\phi) \leftrightarrow (\langle a\rangle^{\#1}\phi')$.

PROOF. For any ALX1 model $M = \langle W, cw, \succ, R_a, V\rangle$, and any $w \in W$,

(A1#1) $\langle a\rangle^{\#1}\bot \leftrightarrow \bot$.

	$M, w \Vdash \langle a\rangle^{\#1}\bot$	
$\Leftrightarrow$	$\exists z(R^{a\#}wz$ and $z \in [\![\bot]\!]_M)$	(Truth condition)
$\Rightarrow$	$\exists z(z \in \emptyset)$	(Meta reasoning)
$\Rightarrow$	**False**	(Meta reasoning)

By propositional logic, $\bot \rightarrow \ \langle a\rangle\bot$. So $\langle a\rangle\bot \leftrightarrow \bot$.

(A2#1) $\langle a\rangle^{\#1}(\phi \vee \psi) \leftrightarrow \langle a\rangle^{\#1}\phi \ \vee \langle a\rangle^{\#1}\psi$.

$M, w \Vdash \langle a \rangle^{\#1}(\phi \vee \psi)$

$\Leftrightarrow$ $\exists z(R^{a\#}wz$ and $(z \in [\![\phi]\!]_M$ or $z \in [\![\psi]\!]_M))$ (Truth condition)
$\Leftrightarrow$ $\exists z(R^{a\#}wz$ and $z \in [\![\phi]\!]_M)$ or $\exists z(R^{a\#}wz$ and $z \in [\![\psi]\!]_M)$ (Meta reasoning)
$\Leftrightarrow$ $M, w \Vdash (\langle a \rangle^{\#1}\phi \ \vee \langle a \rangle^{\#1}\psi)$ (Truth condition)

(A3#1) $\langle a \rangle^{\#1}(\phi \wedge \psi) \rightarrow (\langle a \rangle^{\#1}\phi \ \wedge \langle a \rangle^{\#1}\psi)$.

$M, w \Vdash \langle a \rangle^{\#1}(\phi \wedge \psi)$
$\Leftrightarrow$ $\exists z(R^{a\#}wz$ and $z \in [\![\phi \wedge \psi]\!]_M)$ (Truth condition)
$\Leftrightarrow$ $\exists z(R^{a\#}wz$ and $z \in [\![\phi]\!]_M$ and $z \in [\![\psi]\!]_M)$ (Truth condition)
$\Rightarrow$ $\exists z(R^{a\#}wz$ and $z \in [\![\phi]\!]_M)$ and $\exists z(R^{a\#}wz$ and $z \in [\![\psi]\!]_M)$ (Meta reasoning)
$\Leftrightarrow$ $M, w \Vdash \langle a \rangle^{\#1}\phi$ and $M, w \Vdash \langle a \rangle^{\#1}\psi$ (Truth condition)
$\Leftrightarrow$ $M, w \Vdash (\langle a \rangle^{\#1}\phi \ \wedge \langle a \rangle^{\#1}\psi)$ (Truth condition)

(A#) $\langle a \rangle^{\#1}\phi \rightarrow \langle a \rangle\phi$.

$M, w \Vdash \langle a \rangle^{\#1}\phi$
$\Leftrightarrow$ $\exists z(R^{a\#}wz$ and $z \in [\![\phi]\!]_M)$ (Truth condition)
$\Rightarrow$ $\exists z(R^{a}wz$ and $z \in [\![\phi]\!]_M)$ $(R^{a\#}wz \Rightarrow R^{a}wz)$
$\Leftrightarrow$ $M, w \Vdash \langle a \rangle\phi$ (Truth condition)

The proofs about the inference rules are straightforward. □

Minimal Change Action Focusing on the Closest Worlds

A second possible definition would approach the minimal change via minimally different ϕ-worlds. Call the corresponding minimal change action $\langle a \rangle^{\#2}$. The corresponding truth condition is:

$[\![\langle a \rangle^{\#2}\phi]\!]_M = \{w : \exists w' \in W(w' \in cw(w, [\![\phi]\!]_M)$ and $R^a ww')\}$.

Reconsider the previous example for $\langle a \rangle^{\#2}$. Slamming the door would now get us to worlds where the door is shut and the picture is back on the wall.

Since an additional action is implicit in $\langle a \rangle^{\#2}$, this kind of minimal change appears less intuitive than $\langle a \rangle^{\#1}$. However, $\langle a \rangle^{\#1}$ has a drawback as well: $\langle a \rangle^{\#1}$ will give counterintuitive result if the intersection of accessible worlds and ϕ-worlds is not empty, although the intersection of the closest a-accessible worlds with the ϕ-worlds is.
We can show that the following axioms and inference rules for $\langle a \rangle^{\#2}$ are valid on the

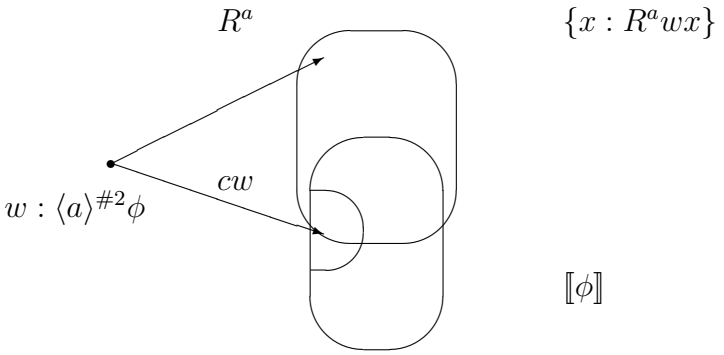


Figure 10.2: Minimal change actions focusing on the closest worlds

class of ALX1 models.

$$\begin{array}{lllll}
\textbf{(}\boldsymbol{A}\textbf{1\#2)}: & \langle a\rangle^{\#2}\bot & \leftrightarrow & \bot. \\
\textbf{(}\boldsymbol{A}\textbf{2\#2)}: & \langle a\rangle^{\#2}\phi & \rightarrow & \langle a\rangle(\langle a\rangle\phi \circ \phi). \\
\textbf{(}\boldsymbol{A}\textbf{\#2)}: & \langle a\rangle^{\#2}\phi & \rightarrow & \langle a\rangle\phi.
\end{array}$$

$$\textbf{(}\boldsymbol{SUBA}\textbf{\#2)}: \ \vdash (\phi \leftrightarrow \phi') \ \Rightarrow \ \vdash (\langle a\rangle^{\#2}\phi) \leftrightarrow (\langle a\rangle^{\#2}\phi').$$

PROOF. For any ALX1 model $M = \langle W, cw, \succ, R_a, V\rangle$, and any $w \in W$,

(A1#2) $\langle a\rangle^{\#2}\bot \leftrightarrow \bot$.

$$\begin{array}{lll}
 & M, w \Vdash \langle a\rangle^{\#2}\bot & \\
\Leftrightarrow & \exists z(R^a wz \text{ and } z \in cw(w, [\![\bot]\!]_M)) & \text{(Truth condition)} \\
\Rightarrow & \exists z(z \in [\![\bot]\!]_M) & \text{(CS1)} \\
\Rightarrow & \exists z(z \in \emptyset) & \text{(Meta reasoning)} \\
\Rightarrow & \textbf{False} & \text{(Meta reasoning)}
\end{array}$$

By propositional logic, $\bot \rightarrow \langle a\rangle\bot$. So $\langle a\rangle^{\#2}\bot \leftrightarrow \bot$.

(A2#2) $\langle a\rangle^{\#2}\phi \rightarrow \langle a\rangle(\langle a\rangle\phi \circ \phi)$.

$$
\begin{array}{lll}
& M, w \Vdash \langle a \rangle^{\#2}\phi & \\
\Leftrightarrow & \exists z(R^a wz \text{ and } (z \in cw(w, [\![\phi]\!]_M)) & \text{(Truth condition)} \\
\Rightarrow & \exists z(R^a wz \text{ and } z \in [\![\phi]\!]_M \text{ and } (z \in cw(w, [\![\phi]\!]_M))) & \text{(CS1)} \\
\Leftrightarrow & \exists z(R^a wz \text{ and } (R^a wz \text{ and } z \in [\![\phi]\!]_M) \text{ and } (z \in cw(w, [\![\phi]\!]_M)) & \text{(Meta reasoning)} \\
\Leftrightarrow & \exists z(R^a wz \text{ and } w \in [\![\langle a \rangle \phi]\!]_M \text{ and } (z \in cw(w, [\![\phi]\!]_M))) & \text{(Truth condition)} \\
\Leftrightarrow & \exists z(R^a wz \text{ and } z \in [\![\langle a \rangle \phi \circ \phi]\!]_M) & \text{(Truth condition)} \\
\Leftrightarrow & M, w \Vdash \langle a \rangle(\langle a \rangle \phi \circ \phi) & \text{(Truth condition)}
\end{array}
$$

(A#2) $\langle a \rangle^{\#2}\phi \rightarrow \langle a \rangle \phi$.

$$
\begin{array}{lll}
& M, w \Vdash \langle a \rangle^{\#2}\phi & \\
\Leftrightarrow & \exists z(R^a wz \text{ and } z \in cw(w, [\![\phi]\!]_M)) & \text{(Truth condition)} \\
\Rightarrow & \exists z(R^a wz \text{ and } z \in [\![\phi]\!]_M & \text{(CS1)} \\
\Leftrightarrow & M, w \Vdash \langle a \rangle \phi & \text{(Truth condition)}
\end{array}
$$

The proof about (SUBA#2) is straightforward. □

Minimal Change Actions Focusing on Both Accessible Worlds and Closest Worlds

As mentioned above, $\langle a \rangle^{\#1}$ has a drawback as well: $\langle a \rangle^{\#1}$ will give counterintuitive result if the intersection of accessible worlds and ϕ-worlds is not empty, although the intersection of the closest a-accessible worlds with the ϕ-worlds is. This could happen, for example, if slamming the door would not shut the door (because of reverberation of the door frame, for instance). To cover this possibility we might want to look at the closest world in the intersection of a-accessible worlds and ϕ-worlds. On this view, the minimal change action is not going to return the empty set if the intersection of the accessible worlds and the closest worlds is not empty. Denote this kind of minimal change by $\langle a \rangle^{\#3}$. The corresponding truth condition is:

$$[\![\langle a \rangle^{\#3}\phi]\!]_M = \{w : \exists w' \in W(w' \in cw(w, \{x : R^a wx\} \cap [\![\phi]\!]_M))\}.$$

But $\langle a \rangle^{\#3}$ cannot be the last word either, because it is leaving the question undecided whether the picture is on the wall or not. In sum, we should avoid a full fledged definition of the closest world function until we can decide this – and possibly other – questions.
We can show that the following propositions and inference rules about the minimal

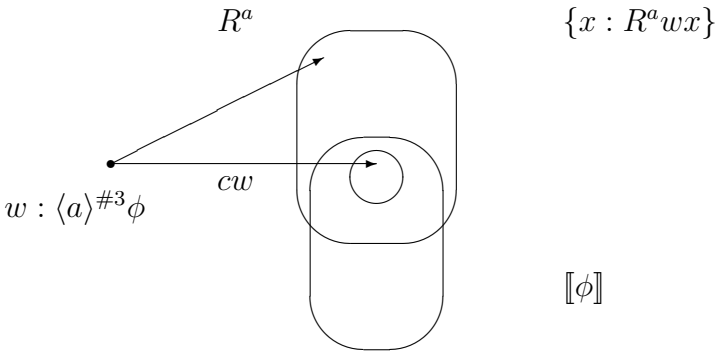


Figure 10.3: Minimal change actions focusing on both accessible worlds and closest worlds

change action are valid on the class of ALX1 models.

$$\begin{array}{lllll}
\textbf{(}\boldsymbol{A1\#3}\textbf{)}: & \langle a\rangle^{\#3}\bot & \leftrightarrow & \bot. \\
\textbf{(}\boldsymbol{A\#3}\textbf{)}: & \langle a\rangle^{\#3}\phi & \rightarrow & \langle a\rangle\phi.
\end{array}$$

$$\textbf{(}\boldsymbol{SUBA\#3}\textbf{)}: \ \vdash(\phi \leftrightarrow \phi') \ \Rightarrow \ \vdash(\langle a\rangle^{\#3}\phi) \leftrightarrow (\langle a\rangle^{\#3}\phi').$$

PROOF. For any ALX1 model $M = \langle W, cw, \succ, R_a, V\rangle$, and any $w \in W$,

(A1#3) $\langle a\rangle^{\#3}\bot \leftrightarrow \bot$.

$$\begin{array}{lll}
 & M, w \Vdash \langle a\rangle^{\#3}\bot & \\
\Leftrightarrow & \exists z(z \in cw(w, \{x : R^a wx\} \cap [\![\bot]\!]_M)) & \text{(Truth condition)} \\
\Rightarrow & \exists z(z \in \{x : R^a wx\} \cap [\![\bot]\!]_M) & \text{(CS1)} \\
\Rightarrow & \exists z(z \in [\![\bot]\!]_M) & \text{(Meta reasoning)} \\
\Rightarrow & \exists z(z \in \emptyset) & \text{(Meta reasoning)} \\
\Rightarrow & \textbf{False} & \text{(Meta reasoning)}
\end{array}$$

By propositional logic, $\bot \rightarrow \langle a\rangle\bot$. So $\langle a\rangle^{\#3}\bot \leftrightarrow \bot$.

(A#3) $\langle a\rangle^{\#3}\phi \rightarrow \langle a\rangle\phi$.

$$
\begin{array}{lll}
 & M, w \Vdash \langle a \rangle^{\#3}\phi & \\
\Leftrightarrow & \exists z(z \in cw(w, \{x : R^a wx\} \cap [\![\phi]\!]_M)) & \text{(Truth condition)} \\
\Rightarrow & \exists z(R^a wz \text{ and } z \in cw(w, \{x : R^a wx\} \cap [\![\phi]\!]_M)) & \text{(Meta reasoning)} \\
\Rightarrow & \exists z(R^a wz \text{ and } z \in \{x : R^a wx\} \cap [\![\phi]\!]_M) & \text{(CS1)} \\
\Rightarrow & \exists z(R^a wz \text{ and } z \in [\![\phi]\!]_M) & \text{(Truth condition)} \\
\Leftrightarrow & M, w \Vdash \langle a \rangle\phi & \text{(Truth condition)}
\end{array}
$$

The proof about (SUBA#3) is straightforward. □

Moreover, if cw satisfies the condition (CSN),

$$(CSN) \quad X \neq \emptyset \Rightarrow cw(w, X) \neq \emptyset.$$

we can show that $\langle a \rangle\phi \leftrightarrow \langle a \rangle^{\#3}\phi$ is valid in the class of ALX1 models. This means that it is not necessary to define such an additional operator, since the action operator $\langle a \rangle$ and the minimal change action $\langle a \rangle^{\#}$ are the same.

10.6.1. Claim. *If the condition (CSN) holds in an ALX1 model M, then $\langle a \rangle\phi \leftrightarrow \langle a \rangle^{\#3}\phi$ is valid on the model M, for any formula ϕ.*

PROOF.
($\Leftarrow$): straightforward from (A#3).
($\Rightarrow$):

$$
\begin{array}{lll}
 & M, w \Vdash \langle a \rangle\phi & \\
\Leftrightarrow & \exists z(R^a wz \text{ and } z \in [\![\phi]\!]_M) & \text{(Truth condition)} \\
\Leftrightarrow & \exists z(z \in \{x : R^a wx\} \text{ and } z \in [\![\phi]\!]_M) & \text{(Meta reasoning)} \\
\Rightarrow & (\{x : R^a wx\} \cap [\![\phi]\!]_M) \neq \emptyset & \text{(Meta reasoning)} \\
\Rightarrow & cw(w, (\{x : R^a wx\} \cap [\![\phi]\!]_M)) \neq \emptyset & \text{(CSN)} \\
\Leftrightarrow & \exists z(z \in cw(w, (\{x : R^a wx\} \cap [\![\phi]\!]_M))) & \text{(Meta reasoning)} \\
\Leftrightarrow & M, w \Vdash \langle a \rangle^{\#3}\phi & \text{(Truth condition)}
\end{array}
$$

□

Chapter 11

ALX3: A Multi-agent ALX Logic

11.1 Introduction

In order to obtain a more powerful version of ALX logic to serve our application, we need a predicate version of ALX logic which contains the first order predicate logic (without equality and function) as one of its subsystem. ALX2 is such an ALX logic. However, it would be more useful to have a multi-agent vesion of ALX logic which extends ALX2. ALX3 is a multi-agent version of first order ALX. ALX3 extends ALX2 by allowing multi-agent formulas and the occurrence of action and agent variables in first order formulas. As mentioned in the last chapter, ALX3 contains a logic of conditional instead of update, since we want to study this alternative. Furthermore, ALX3 also contains a logic of belief as its subsystem. Actually, ALX2 is only a subsystem of ALX3. In this chapter, we focus on ALX3 logic. For more details about ALX2, see [Huang, Masuch&Pólos 1993].

11.2 Formal Syntax and Semantics

11.2.1 Syntax

The language has the following primitive symbols:

(1) For each natural number $n(\geq 1)$, a countable set of n-place predicate letters, PRE_n, which we write as $p_i, p_j, ...$

(2.1) A countable set of regular variables, $RVAR$, which we write as $x, x_1, y, z, ...$

(2.2) A countable set of action variables, $AVAR$, which we write as $a, a_1, b, ...$

(2.3) A countable set of agent variables, $AGVAR$, which we write as $i, i_1, j, ...$

(3.1) A countable set of regular constants, $RCON$, which we write as $c, c_1, c_2, ...$

(3.2) A countable set of actions constants, $ACON$, which we write as $ac, ac_1, ac_2, ...$

(3.3) A countable set of agent constants, $AGCON$, which we write as $ag, ag_1, ag_2, ...$

(4) The symbols $\neg$(negation), $\wedge$(conjunction), $\boldsymbol{B}$(belief), $\exists$ (existential quantifier), $\boldsymbol{P}$(preference), $\rightsquigarrow$(conditional), ;(sequence), $\cup$(choice), and parentheses: $\langle, \rangle$, (, and).

We assume that the above symbol sets are disjoint so their intersections are empty.

11.2.1. Definition. (Variable) *The set of variables VAR is defined as follows:*

$$VAR = RVAR \cup AVAR \cup AGVAR.$$

11.2.2. Definition. (Constant) *The set of constants CON is defined as follows:*

$$CON = RCON \cup ACON \cup AGCON.$$

11.2.3. Definition. (Term) *The set of terms $TERM$ is defined as follows:*

$$TERM = VAR \cup CON.$$

11.2.4. Definition. (Action term) *The set of action terms $ATERM$ is defined as follows:*

$$ATERM = AVAR \cup ACON.$$

11.2.5. Definition. (Agent term) *The set of agent terms $AGTERM$ is defined as follows:*

$$AGTERM = AGVAR \cup AGCON.$$

In the following, we use $t, t_1, ...,$ to denote terms, $a, a_1, ...,$ to denote action terms, $i, j, ...,$ to denote agent terms if that does not cause any ambiguity.

11.2.6. Definition. (Atom) *The set of atomic formulae $ATOM$ is defined as follows:*

$$ATOM =_{df} \{p(t_1, t_2, ..., t_n) \ : \ p \in PRE_n, t_1, t_2, ..., t_n \in TERM\}.$$

11.2.7. Definition. (Action) *The set of action expression $ACTION$ is defined recursively as follows:*

- $a \in ATERM, i \in AGTERM \Rightarrow a_i \in ACTION.$
- $a, b \in ACTION \Rightarrow (a; b), (a \cup b) \in ACTION.$

11.2.8. Definition. (Formula) *The set of formulae FML is defined recursively as follows:*

- $ATOM \subseteq FML.$
- $\phi \in FML \Rightarrow \neg\phi \in FML.$
- $\phi, \psi \in FML \Rightarrow (\phi \wedge \psi) \in FML.$

- $\phi \in FML, x \in VAR \Rightarrow (\exists x\phi) \in FML.$
- $\phi \in FML, a \in ACTION \Rightarrow (\langle a \rangle \phi) \in FML.$
- $\phi, \psi \in FML \Rightarrow (\phi \rightsquigarrow \psi) \in FML.$
- $\phi, \psi \in FML, i \in AGTERM \Rightarrow (\phi \boldsymbol{P}_i \psi) \in FML.$
- $\phi \in FML, i \in AGTERM \Rightarrow \boldsymbol{B}_i \phi \in FML.$

Define $\bot$ as $\phi \wedge \neg\phi$ for an arbitrary ϕ, and $[a]\phi$ as $\neg\langle a \rangle \neg\phi$. Define the boolean connectives $\{\vee, \rightarrow, \leftrightarrow\}$, and the truth constant $\top$ from the given boolean connectives in the usual way. $\forall x(\varphi)$ is defined as $\neg\exists x(\neg\varphi)$.

In the following, lower case Greek letter ϕ, ψ, ..., with or without subscript or superscript denote formulae; lower case Latin symbols a, b with or without subscript or superscript denote actions; lower case Latin symbols t with or without subscript or superscript denote terms; lower case Latin symbols x, y, z with or without subscript or superscript denote variables if they cannot cause any ambiguity.

11.2.2 Semantics

11.2.9. Definition. (ALX3 model) *We call*
$M = \langle O, PA, AGENT, W, cw, \succ, \mathcal{R}, \mathcal{B}, I \rangle$ *an ALX3 model, if*

- O *is a set of objects,*
- PA *is a set of primitive actions,*
- $AGENT$ *is a set of agents,*
- W *is a set of possible worlds,*
- $cw : W \times \mathcal{P}(W) \hookrightarrow \mathcal{P}(W)$ *is a closest world function,*
- $\succ$: $AGENT \rightarrow \mathcal{P}(\mathcal{P}(W) \times \mathcal{P}(W))$ *is a function which assigns a comparison relation for preferences to each agent,*
- $\mathcal{R} : AGENT \times PRIMITIVE\text{-}ACTION \rightarrow \mathcal{P}(W \times W)$ *is a function which assigns an accessibility relation to each agent and each primitive action,*
- $\mathcal{B} : AGENT \rightarrow \mathcal{P}(W \times W)$ *is a function that assigns an accessibility relation for belief to each agent,*
- I *is a pair* $\langle I_P, I_C \rangle$,
  where I_P *is a predicate interpretation function which assigns to each n-place predicate letter* $p \in PRE_n$ *and each world* $w \in W$ *a set of n tuples* $\langle u_1, ..., u_n \rangle$, *where each of* $u_1, .., u_n$ *is in* $D = O \cup PA \cup AGENT$, *called a domain, and*
  I_C *is a constant interpretation function which assigns to each regular constants* $c \in RCON$ *an object* $d \in O$, *assigns to each action constant* $ac \in ACON$ *a primitive action* $a_p \in PA$, *and assigns to each agent constant* $g \in AGCON$ *an agent* $a_g \in AGENT$.

and if the model satisfies the following conditions:

(i) the closest world function cw satisfies (CS1)-(CS5).
(ii)the comparison relation (for each agent) satisfies the following conditions:

For each agent $i \in AGENT$,

$(\boldsymbol{NORM})$: $(\emptyset \not\succ_i X), (X \not\succ_i \emptyset)$.
where $\succ_i = \succ (i)$.

$(\boldsymbol{TRAN})$: $cw(w, X \cap \overline{Y}) \succ_i cw(w, Y \cap \overline{X})$ *and* $cw(w, y \cap \overline{Z}) \succ_i cw(w, Z \cap \overline{Y})$ $\Rightarrow cw(w, X \cap \overline{Z}) \succ_i cw(w, Z \cap \overline{X})$,

where $\overline{X} = W - X$.

(iii) the accessibility relation (for each agent) is serial and transitive, namely,

$(\boldsymbol{SEB})$: $\forall w \exists w'(\langle w, w' \rangle \in \mathcal{B}_i)$,

where $\mathcal{B}_i = \mathcal{B}(i)$.

$(\boldsymbol{TRB})$: $\langle w, w' \rangle \in \mathcal{B}_i$ *and* $\langle w', w'' \rangle \in \mathcal{B}_i \Rightarrow \langle w, w'' \rangle \in \mathcal{B}_i$

As argued in the chapter 8, (CS1)-(CS5) are standard constraints for the closest world function in semantic model of conditionals. (NORM) and (TRAN) constrain the same semantics for preference relation as we do in ALX1. (SEB) and (TRB) make the operator as a real belief operator, namely, a weak S4 one.

11.2.10. Definition. (Valuation of variables) *A valuation of variables* v *in the domain* D *of an ALX3 model* M *is a mapping which assigns to each variable* $x \in VAR$ *an element* $d \in D$ *such that* $v(x) \in O$, $v(a) \in PA$, *and* $v(i) \in AGENT$ *for any* $x \in RVAR, a \in AVAR$, *and* $i \in AGVAR$.

11.2.11. Definition. (Valuation of terms) *For an ALX3 model* $M = \langle O, PA, AGENT, W, cw, \succ, \mathcal{R}, \mathcal{B}, I \rangle$ *and a valuation of variables* v, *a valuation of terms* v_I *is a function which assigns to each term* $t \in TERM$ *an element in the domain* D, *which is defined as follows:*

$$t \in CON \Rightarrow v_I(t) = I_C(t);$$

$$t \in VAR \Rightarrow v_I(t) = v(t).$$

Suppose that v is a valuation of variables, d is an element of domain, and x is a variable. We use the notation $v(d/x)$ to denote the valuation of variables which assigns the same values to the variables as does v except that it assigns the value d to x. Moreover, we use the notation V_D to denote the set of valuations of variables in the domain D.

11.2.12. Definition. (Accessibility relations for actions) *Define an accessibility relation* $R^{a'}$ *in a model* $M = \langle O, PA, AGENT, W, cw, \succ, \mathcal{R}, \mathcal{B}, I \rangle$ *and a valuation* v *for each action* $a' \in ACTION$ *as follows.*

- $a \in ATERM, i \in AGTERM \Rightarrow R^{a_i} = \mathcal{R}(v_I(a), v_I(i))$,

- $a, b \in ACTION \Rightarrow R^{(a;b)} = R^a \circ R^b = \{\langle w, w'\rangle \in W \times W : (\exists w_1 \in W)(R^a w w_1 \text{ and } R^b w_1 w')\}$,
- $a, b \in ACTION \Rightarrow R^{(a \cup b)} = R^a \cup R^b$.

11.2.13. Definition. (Meaning function) *Let FML be as above and let $M = \langle O, PA, AGENT, W, cw, \succ, \mathcal{R}, \mathcal{B}, I\rangle$ be an ALX3 model and let v be a valuation of variables in the domain D. The meaning function $[\![\ ]\!]^v_M : FML \to \mathcal{P}(\mathcal{W})$ is defined as follows:*

$$
\begin{array}{lcl}
[\![p(t_1, ..., t_n)]\!]^v_M & = & \{w \in W : \langle v_I(t_1), v_I(t_2), ..., v_I(t_n)\rangle \in I_P(p, w)\} \quad \text{where } p \in PRE_n. \\
[\![\neg\phi]\!]^v_M & = & W \setminus [\![\phi]\!]^v_M. \\
[\![\phi \wedge \psi]\!]^v_M & = & [\![\phi]\!]^v_M \cap [\![\psi]\!]^v_M. \\
[\![\exists x\phi]\!]^v_M & = & \{w \in W : (\exists d \in D)(w \in [\![\phi]\!]^{v(d/x)}_M)\}. \\
[\![\langle a\rangle\phi]\!]^v_M & = & \{w \in W : (\exists w' \in W)(R^a w w' \text{ and } w' \in [\![\phi]\!]^v_M)\}. \\
[\![\phi \rightsquigarrow \psi]\!]^v_M & = & \{w \in W : cw(w, [\![\phi]\!]^v_M) \subseteq [\![\psi]\!]^v_M\}. \\
[\![\phi \boldsymbol{P}_i \psi]\!]^v_M & = & \{w \in W : cw(w, [\![\phi \wedge \neg\psi]\!]^v_M) \succ_{v_I(i)} cw(w, [\![\psi \wedge \neg\phi]\!]^v_M)\}. \\
[\![\boldsymbol{B}_i\phi]\!]^v_M & = & \{w \in W : (\forall w')(\langle w, w'\rangle \in \mathcal{B}_{v_I(i)} \Rightarrow w' \in [\![\phi]\!]^v_M)\}.
\end{array}
$$

As argued before, those interpretations are standard.[1] The forcing and satisfiable relations are defined as usual.

11.2.14. Definition. (ALX3 inference system) *Let ALX3S be the following set of axioms and rules of inference.*

$$
\begin{array}{lll}
(\boldsymbol{BA}) : & \text{all tautologies of the first order logic} & \\
 & & \\
(\boldsymbol{A1}) : & \langle a\rangle\bot & \leftrightarrow \ \bot. \\
(\boldsymbol{A2}) : & \langle a\rangle(\phi \vee \psi) & \leftrightarrow \ \langle a\rangle\phi \ \vee \langle a\rangle\psi. \\
(\boldsymbol{A3}) : & \langle a; b\rangle\phi & \leftrightarrow \ \langle a\rangle\langle b\rangle\phi. \\
(\boldsymbol{A4}) : & \langle a \cup b\rangle\phi & \leftrightarrow \ \langle a\rangle\phi \vee \langle b\rangle\phi. \\
(\boldsymbol{AU}) : & [a]\forall x\phi & \rightarrow \ \forall x[a]\phi. \\
 & & \\
(\boldsymbol{ID}) : & \psi \rightsquigarrow \psi. & \\
(\boldsymbol{MPC}) : & (\psi \rightsquigarrow \phi) & \rightarrow \ (\psi \rightarrow \phi). \\
(\boldsymbol{CC}) : & (\psi \rightsquigarrow \phi) \wedge (\psi \rightsquigarrow \phi') & \rightarrow \ \psi \rightsquigarrow (\phi \wedge \phi'). \\
(\boldsymbol{MOD}) : & (\neg\psi \rightsquigarrow \psi) & \rightarrow \ (\phi \rightsquigarrow \psi). \\
(\boldsymbol{CSO}) : & ((\psi \rightsquigarrow \phi) \wedge (\phi \rightsquigarrow \psi)) & \rightarrow \ ((\psi \rightsquigarrow \chi) \leftrightarrow (\phi \rightsquigarrow \chi)). \\
(\boldsymbol{CV}) : & ((\psi \rightsquigarrow \phi) \wedge \neg(\psi \rightsquigarrow \neg\chi)) & \rightarrow \ ((\psi \wedge \chi) \rightsquigarrow \phi). \\
(\boldsymbol{CS}) : & (\psi \wedge \phi) & \rightarrow \ (\psi \rightsquigarrow \phi).
\end{array}
$$

[1] For the interpretation of the existential quantifier, in fact the logic requires an interpretation for the three types of variables involves, by the definition of the valuation of variables.

$$
\begin{array}{lll}
(\boldsymbol{CEP}): & \phi \boldsymbol{P}_i \psi & \leftrightarrow (\phi \wedge \neg\psi)\boldsymbol{P}_i(\neg\phi \wedge \psi). \\
(\boldsymbol{N}): & \neg(\bot \boldsymbol{P}_i \phi), & \neg(\phi \boldsymbol{P}_i \bot). \\
(\boldsymbol{TR}): & (\phi \boldsymbol{P}_i \psi) \wedge (\psi \boldsymbol{P}_i \chi) & \rightarrow (\phi \boldsymbol{P}_i \chi). \\
 & & \\
(\boldsymbol{PC}): & (\phi \boldsymbol{P}_i \psi) & \rightarrow \neg((\phi \wedge \neg\psi) \rightsquigarrow \neg(\phi \wedge \neg\psi)) \wedge \\
 & & \quad \neg((\psi \wedge \neg\phi) \rightsquigarrow \neg(\psi \wedge \neg\phi)). \\
 & & \\
(\boldsymbol{KB}): & \boldsymbol{B}_i\phi \wedge \boldsymbol{B}_i(\phi \rightarrow \psi) & \rightarrow \boldsymbol{B}_i\psi. \\
(\boldsymbol{DB}): & \neg \boldsymbol{B}_i \bot. & \\
(\boldsymbol{4B}): & \boldsymbol{B}_i\phi & \rightarrow \boldsymbol{B}_i\boldsymbol{B}_i\phi. \\
 & & \\
(\boldsymbol{BFB}): & \forall x \boldsymbol{B}_i \phi & \rightarrow \boldsymbol{B}_i \forall x \phi.
\end{array}
$$

$$
\begin{array}{lll}
(\boldsymbol{MP}): & \vdash \phi \ \& \ \vdash \phi \rightarrow \psi & \Rightarrow \ \vdash \psi. \\
(\boldsymbol{G}): & \vdash \phi & \Rightarrow \ \vdash \forall x \phi. \\
(\boldsymbol{NECA}): & \vdash \phi & \Rightarrow \ \vdash [a]\phi. \\
(\boldsymbol{NECB}): & \vdash \phi & \Rightarrow \ \vdash \boldsymbol{B}_i\phi. \\
(\boldsymbol{MONA}): & \vdash \langle a \rangle \phi \ \& \ \vdash \phi \rightarrow \psi & \Rightarrow \ \vdash \langle a \rangle \psi. \\
(\boldsymbol{MONC}): & \vdash \phi \rightsquigarrow \psi \ \& \ \vdash \psi \rightarrow \psi' & \Rightarrow \ \vdash \phi \rightsquigarrow \psi'. \\
(\boldsymbol{SUBA}): & \vdash (\phi \leftrightarrow \phi') & \Rightarrow \ \vdash (\langle a \rangle \phi) \leftrightarrow (\langle a \rangle \phi'). \\
(\boldsymbol{SUBC}): & \vdash (\phi \leftrightarrow \phi') \ \& \ \vdash (\psi \leftrightarrow \psi') & \Rightarrow \ \vdash (\phi \rightsquigarrow \psi) \leftrightarrow (\phi' \rightsquigarrow \psi'). \\
(\boldsymbol{SUBP}): & \vdash (\phi \leftrightarrow \phi') \ \& \ \vdash (\psi \leftrightarrow \psi') & \Rightarrow \ \vdash (\phi \boldsymbol{P}_i \psi) \leftrightarrow (\phi' \boldsymbol{P}_i \psi').
\end{array}
$$

Most axioms are straightforward. As usual, we have the tautologies (BA) and the generalization rule (G). Since ALX3 is a normal modal logic with respect to the action operator, the absurdum is not true anywhere, so it is not accessible (A1). The action modalities behave as usual, so they distribute over disjunction both ways, but over conjunction only in one direction (A2). The axiom (PC) says that if an agent i prefers ϕ to ψ, then both $\phi \wedge \neg\psi$ and $\psi \wedge \neg\phi$ are possible. The axioms (BA), (KB), (DB), (4B), and the inference rules (MP), (NECB) turn the belief operation into a weak $S4$ system.

We have the modus ponens and the necessitation rule for the universal action modality (NECA), and monotonicity for the existential action modality. For conditionals, we have *right* monotonicity, but not left monotonicity. Logically equivalent propositions are substitutional in action- conditional- and preference formulae (SUBA), (SUBC), (SUBP). Note that we are *not* having monotonicity for preferences. Because of this, we are able to avoid the counterintuitive deductive closure of goals.

11.3 Formal Properties of ALX3

11.3.1 Soundness

11.3.1. Proposition. (Soundness of ALX3S) *ALX3S is sound.*

PROOF. (A4) $\langle a \cup b\rangle\phi \leftrightarrow \langle a\rangle\phi \vee \langle b\rangle\phi$.

$$
\begin{array}{lll}
& M, w, v\|\!-\langle a \cup b\rangle\phi & \\
\Leftrightarrow & \exists z(\langle w, z\rangle \in R^a \cup R^b \text{ and } z \in [\![\phi]\!]^v_M) & \text{(Truth condition)} \\
\Leftrightarrow & \exists z(\langle w, z\rangle \in R^a \text{ and } z \in [\![\phi]\!]^v_M) \text{ or } \exists z(\langle w, z\rangle \in R^b \text{ and } z \in [\![\phi]\!]^v_M) & \text{(Meta reasoning)} \\
\Leftrightarrow & M, w, v\|\!-\langle a\rangle\phi \vee \langle b\rangle\phi & \text{(Meta reasoning)}
\end{array}
$$

(AU) $[a]\forall x\phi \rightarrow \forall x[a]\phi$.

$$
\begin{array}{lll}
& M, w, v\|\!-[a]\forall x\phi & \\
\Leftrightarrow & (\forall w' \in W)(R^a w w' \Rightarrow M, w, v\|\!-\forall x\phi) & \text{(Truth condition)} \\
\Leftrightarrow & (\forall w' \in W)(R^a w w' \Rightarrow (\forall d \in D)(M, w, v(d/x)\|\!-\phi)) & \text{(Truth condition)} \\
\Leftrightarrow & (\forall w' \in W)(\neg R^a w w' \text{ or } (\forall d \in D)(M, w, v(d/x)\|\!-\phi)) & \text{(Meta reasoning)} \\
\Rightarrow & (\forall w' \in W)(\forall d \in D)(\neg R^a w w' \text{ or } (M, w, v(d/x)\|\!-\phi)) & \text{(Meta reasoning)} \\
\Leftrightarrow & (\forall d \in D)(\forall w' \in W)(R^a w w' \Rightarrow M, w, v(d/x)\|\!-\phi) & \text{(Meta reasoning)} \\
\Leftrightarrow & (\forall d \in D)(M, w, v(d/x)\|\!-[a]\phi) & \text{(Truth condition)} \\
\Leftrightarrow & M, w, v\|\!-\forall x[a]\phi &
\end{array}
$$

(BFB) $\forall x \boldsymbol{B}_i\phi \rightarrow \boldsymbol{B}_i\forall x\phi$.

$$
\begin{array}{lll}
& M, w, v\|\!-\forall x \boldsymbol{B}_i\phi & \\
\Rightarrow & (\forall d)(M, w, v(d/x)\|\!-\boldsymbol{B}_i\phi) & \text{(Truth condition)} \\
\Rightarrow & (\forall d)(\forall w')(\mathcal{B}(v_I(i))ww' \Rightarrow M, w', v(d/x)\|\!-\phi) & \text{(Truth condition)} \\
\Rightarrow & (\forall w')(\mathcal{B}(v_I(i))ww' \Rightarrow (\forall d)M, w', v(d/x)\|\!-\phi) & \text{(Meta reasoning)} \\
\Rightarrow & (\forall w')(\mathcal{B}(v_I(i))ww' \Rightarrow M, w', v\|\!-\forall x\phi) & \text{(Truth condition)} \\
\Rightarrow & M, w, v\|\!-\boldsymbol{B}_i\forall x\phi & \text{(Truth condition)}
\end{array}
$$

The proofs for other axioms and inference rules are straightforward from the definitions and the former proofs. □

11.3.2 More Properties about Action Operators

$$
\begin{array}{lllll}
\textbf{(a1)}: & \langle a\rangle(\phi \wedge \psi) & \rightarrow & \langle a\rangle\phi \wedge \langle a\rangle\psi. \\
\textbf{(a2)}: & \langle a \cup a\rangle\phi & \leftrightarrow & \langle a\rangle\phi. \\
\textbf{(a3)}: & \langle (a \cup b) \cup c\rangle\phi & \leftrightarrow & \langle a \cup (b \cup c)\rangle\phi. \\
\textbf{(a4)}: & \langle a \cup b\rangle\phi & \leftrightarrow & \langle b \cup a\rangle\phi. \\
\textbf{(a5)}: & \langle a; (b \cup c)\rangle\phi & \leftrightarrow & \langle (a; b) \cup (a; c)\rangle\phi. \\
\textbf{(a6)}: & \langle (a \cup b); c\rangle\phi & \leftrightarrow & \langle (a; b)) \cup (a; c)\rangle\phi.
\end{array}
$$

11.3.1. LEMMA. *(a1)-(a6) are theorems of ALX3S.*[2]

[2]Note that the part (a5) is a well known debatable equivalence from the perspective of process theory. The problem involves the precise timing of choices made by agents.

PROOF.
(a1) $\langle a\rangle(\phi \wedge \psi) \rightarrow \langle a\rangle\phi \wedge \langle a\rangle\psi$.

$$\begin{array}{lll} & \vdash \langle a\rangle(\phi \wedge \psi) & \\ \Rightarrow & \vdash \langle a\rangle\phi \text{ and } \vdash \langle a\rangle\psi & \text{(MONA)} \\ \Leftrightarrow & \vdash \langle a\rangle\phi \wedge \langle a\rangle\psi & \text{(Truth condition)} \end{array}$$

(a2) $\langle a \cup a\rangle\phi \leftrightarrow \langle a\rangle\phi$.

$$\begin{array}{lll} & \vdash \langle a \cup a\rangle\phi & \\ \Leftrightarrow & \vdash \langle a\rangle\phi \vee \langle a\rangle\phi & \text{(A2)} \\ \Leftrightarrow & \vdash \langle a\rangle\phi & \text{(Meta reasoning)} \end{array}$$

(a3) $\langle (a \cup b) \cup c\rangle\phi \leftrightarrow \langle a \cup (b \cup c)\rangle\phi$.

$$\begin{array}{lll} & \vdash \langle (a \cup b) \cup c\rangle\phi & \\ \Rightarrow & \vdash \langle a\rangle\phi \vee \langle b\rangle\phi \vee \langle c\rangle\phi & \text{(A2)} \\ \Leftrightarrow & \vdash \langle a \cup (b \cup c)\rangle\phi & \text{(Truth condition)} \end{array}$$

(a4) $\langle a \cup b\rangle\phi \leftrightarrow \langle b \cup a\rangle\phi$ is straigtforward.
(a5) $\langle a; (b \cup c)\rangle\phi \leftrightarrow \langle (a; b) \cup (a; c)\rangle\phi$.

$$\begin{array}{lll} & \vdash \langle a; (b \cup c)\rangle\phi & \\ \Leftrightarrow & \vdash \langle a\rangle\langle b \cup c\rangle\phi & \text{(A3)} \\ \Leftrightarrow & \vdash \langle a\rangle(\langle b\rangle\phi \vee \langle c\rangle\phi) & \text{(A4)} \\ \Leftrightarrow & \vdash (\langle a\rangle\langle b\rangle\phi) \vee (\langle a\rangle\langle c\rangle\phi) & \text{(A2)} \\ \Leftrightarrow & \vdash (\langle a; b\rangle\phi) \vee (\langle a; c\rangle\phi) & \text{(A3)} \\ \Leftrightarrow & \vdash \langle (a; b) \cup (a; c)\rangle\phi & \text{(A4)} \end{array}$$

(a6) $\langle (a \cup b); c\rangle\phi \leftrightarrow \langle (a; b)) \cup (a; c)\rangle\phi$ is similar to (a5). □

11.4 Completeness

First of all, we note that ALX3 combines a predicate dynamic logic, a preference logic (with conditional), and a doxastic logic(i.e., a belief logic) together; all of those semantics components are independent in the combination; and all of the sub-logics are complete with respect to their own semantic components. Therefore, it is reasonable to assume that the completeness can be transferred into the combination. In [Fine&Schurz 1992], Kit Fine and Gerhard Schurz have proved the completeness for the combination of unary modalities transfers. It would be logically more elegant if

we can prove the completeness of ALX3 by a general theorem for independent combination of logics. However, in ALX3, we have a difficulty with the treatment of the binary modalities, i.e., the conditional and preference. Moreover, the two sorts structure of dynamic logic also causes more difficulties so that we cannot easily construct a hanging function like Kit Fine and Gerhard Schurz did in [Fine&Schurz 1992]. Therefore, we have to use the ordinary approach, namely the Henkin approach, to prove the completeness.

We say a formula φ is a sentence if no free variable occurs in φ. We will construct a canonical model M_c where the possible worlds are maximal consistent sets. Furthermore, we will show that for any sentence χ,

(1) $\chi \in w \Leftrightarrow M_c, w, v \| \!- \chi$ for some valuation v, (truth lemma).
(2) M_c is an ALX3 model.

11.4.1. LEMMA. (Action lemma) *For any maximal consistent set w, and any action $a \in ACTION$,*
($\langle a \rangle \phi \in w \Rightarrow$ There exists a maximal consistent set w' such that $\phi \in w'$ and for any $\psi \in w' (\langle a \rangle \psi \in w)$

PROOF. Note that (A1) and (A2) are axioms of ALX3S, and (MONA) is an inference rule of ALX3S. This suffices to show that. Similar to the proof for ALX1, the action lemma holds. Here we do not want to go into details. □

11.4.2. LEMMA. (Conditional lemma) *For any maximal consistent set w, and any sentence ϕ, χ,*
if $\neg(\phi \rightsquigarrow \chi) \in w$, then there exists a maximal consistent set w' such that
(i) $\neg\chi \in w'$, and
(ii) $((\phi \rightsquigarrow \psi) \in w \Rightarrow \psi \in w')$, for any ψ.

PROOF. Suppose that $\neg(\phi \rightsquigarrow \chi) \in w$. We construct a set w" as follows:

$$w" = \{\psi : (\phi \rightsquigarrow \psi) \in w\} \cup \{\neg\chi\}.$$

We claim that w" is consistent, since if that is not the case, then there exist $\psi_1, ..., \psi_n$ such that $\phi \rightsquigarrow \psi_1, ..., \phi \rightsquigarrow \psi_n \in w$ and $\psi_1 \wedge ... \psi_n \wedge \neg\chi$ is inconsistent. Consequently, $(\psi_1 \wedge ... \wedge \psi_n) \rightarrow \chi$ is provable. However, from $(\phi \rightsquigarrow \psi_1), ..., (\phi \rightsquigarrow \psi_n) \in w$, we have $\phi \rightsquigarrow (\psi_1 \wedge ... \wedge \psi_n) \in w$. So, by the monotonicity, we have $\phi \rightsquigarrow \chi \in w$, which contradicts $\neg(\phi \rightsquigarrow \chi) \in w$.

Extend w" into a maximal consistent set w'. Then, it is easy to see that w' is

the maximal consistent set we want. □

11.4.3. LEMMA. (Belief lemma) *For any maximal consistent set w, and any sentence ϕ, if $\neg(\boldsymbol{B}_i\phi) \in w$, then there exists a maximal consistent set w' such that*
(i) $\neg\phi \in w'$, and
(ii) $((\boldsymbol{B}_i\psi) \in w \Rightarrow \psi \in w')$, for any ψ.

PROOF. Suppose that $\neg\boldsymbol{B}_i\phi \in w$. We construct a set w" as follows:

$$w" = \{\psi : \boldsymbol{B}_i\psi \in w\} \cup \{\neg\phi\}.$$

Similar to the last lemma. We have the proof. □

11.4.1. THEOREM. (Completeness of ALX3S) *ALX3S is complete for the class of ALX3 models.*

PROOF. We construct a canonical model $M_c = \langle O, PA, AGENT, W, cw, \succ, \mathcal{R}, \mathcal{B}, I\rangle$ as follows:

$O \stackrel{\text{def}}{=} RCON$,
$PA \stackrel{\text{def}}{=} ACON$,
$AGENT \stackrel{\text{def}}{=} AGCON$,
$W \stackrel{\text{def}}{=} \{w : w \text{ is a maximal consistent set}\}$,

Take a valuation of variable v in the domain $D = O \cup PA \cup AGENT$.

Define the constant interpretation function as follows:
$I_C(c) = c$, for any $c \in CON$,

So, $v_I(c) = c$ for any $c \in CON$,
$v_I(x) = v(x)$ for any $x \in VAR$.

Furthermore, define the predicate interpretation function as follows:
$I_P(p, w) = \{\langle v_I(t_1), ..., v_I(t_n)\rangle : p(t_1, ..., t_n) \in w\}$.

Moreover, we define,

$w \in cw(w', [\![\phi]\!]^v_{M_c})$ iff $\forall\psi((\phi \rightsquigarrow \psi) \in w' \Rightarrow \psi \in w)$.

For each $ag \in AGENT$, we define:

$cw(w, [\![\phi \wedge \neg\psi]\!]^v_{M_c}) \succ_{ag} cw(w, [\![\psi \wedge \neg\phi]\!]^v_{M_c})$ iff $\phi\mathbf{P}_{ag}\psi \in w$,

$\langle w, w' \rangle \in \mathcal{B}(ag)$ iff $\forall \psi(\boldsymbol{B}_{ag}\psi \in w \Rightarrow \psi \in w')$.

For each $ag \in AGENT$, each $ac \in PRIMITIVE\text{-}ACTION$, we define:

$\langle w, w' \rangle \in \mathcal{R}(ag, ac)$ iff $\forall \psi(\psi \in w' \Rightarrow \langle ac_{ag} \rangle \psi \in w)$.

FACT 1(Action expansion property). For the canonical model M_c, the valuation of variable v, and any action a' in which no free variable occurs,
$\langle w, w' \rangle \in R^{a'}$ iff $\forall \rho(\rho \in w' \Rightarrow \langle\!\langle a' \rangle\!\rangle \rho \in w)$.
PROOF. We prove the fact by induction on the complexity of a'.

Case 1. $a' = ac_{ag}$,

$$\begin{array}{lll} & \langle w, w' \rangle \in R^{a'} & \\ \Leftrightarrow & \langle w, w' \rangle \in \mathcal{R}(v_I(ag), v_I(ac)) & \text{(Definition of } R\text{)} \\ \Leftrightarrow & \forall \rho(\rho \in w' \Rightarrow \langle\!\langle a' \rangle\!\rangle \rho \in w) & \text{(Definition of } \mathcal{R}\text{)} \end{array}$$

Case 2 $a' = (a; b)$,
($\Rightarrow$)

$$\begin{array}{lll} & \langle w, w' \rangle \in R^{a;b} & \\ \Leftrightarrow & \langle w, w' \rangle \in R^a \circ R^b & \text{(Definition of } R\text{)} \\ \Leftrightarrow & \exists w"(\langle w, w" \rangle \in R^a \text{ and } \langle w", w' \rangle \in R^b) & \text{(Definition of } \circ\text{)} \\ \Leftrightarrow & \exists w"(\forall \rho(\rho \in w" \Rightarrow \langle a \rangle \rho \in w) \text{ and } \forall \rho(\rho \in w' \Rightarrow \langle b \rangle \rho \in w")) & \text{(Induction hypothesis)} \\ \Rightarrow & \forall \rho(\rho \in w' \Rightarrow \langle a \rangle \langle b \rangle \rho \in w) & \text{(Meta reasoning)} \\ \Leftrightarrow & \forall \rho(\rho \in w' \Rightarrow \langle a; b \rangle \rho \in w) & \text{(A3)} \end{array}$$

($\Leftarrow$)
$\forall \rho(\rho \in w' \Rightarrow \langle a; b \rangle \rho \in w)$
$\Rightarrow \forall \rho(\rho \in w' \Rightarrow \langle a \rangle \langle b \rangle \rho \in w)$.

Let $w_a = \{\neg\varphi : \neg\langle a \rangle \varphi \in w\}$.
Let $w_b = \{\langle b \rangle \varphi : \varphi \in w'\}$. Let $w_0 = w_a \cup w_b$.

We claim that w_0 is consistent. Since if w_0 is inconsistent, then there exist $\phi_1, ..., \phi_n \in w'$ and $\neg\langle a \rangle \phi'_1, ..., \neg\langle a \rangle \phi'_m \in w$ such that
$\vdash \langle b \rangle \phi_1 \wedge ... \wedge \langle b \rangle \phi_n \wedge \neg\phi'_1 \wedge ... \wedge \neg\phi'_m \rightarrow \bot$. However,
$\vdash \langle b \rangle \phi_1 \wedge ... \wedge \langle b \rangle \phi_n \wedge \neg\phi'_1 \wedge ... \wedge \neg\phi'_m \rightarrow \bot$
$\Rightarrow \vdash \neg(\langle b \rangle \phi \wedge ... \wedge \langle b \rangle \phi_n) \vee (\phi'_1 \vee ... \vee \phi'_m)$
$\Rightarrow \vdash (\langle b \rangle \phi_1 \wedge ... \wedge \langle b \rangle \phi_n) \rightarrow (\phi'_1 \vee ... \vee \phi'_m)$.

On the other hand,

$$
\begin{array}{lll}
 & (\phi_1 \wedge ... \wedge \phi_n) \in w', (\neg\langle a\rangle\phi'_1 \wedge ... \wedge \neg\langle a\rangle\phi'_m) \in w & \\
\Rightarrow & \langle a\rangle\langle b\rangle(\phi_1 \wedge ... \wedge \phi_n) \in w, \neg\langle a\rangle\phi'_1 \wedge ... \wedge \neg\langle a\rangle\phi'_m \in w & (\forall\rho(\rho \in w' \Rightarrow \langle a\rangle\langle b\rangle\rho \in w)) \\
\Rightarrow & \langle a\rangle(\langle b\rangle\phi_1 \wedge ... \wedge \langle b\rangle\phi_n) \in w, \neg\langle a\rangle\phi'_1 \wedge ... \wedge \neg\langle a\rangle\phi'_m \in w & \text{(a1)} \\
\Rightarrow & \langle a\rangle(\phi'_1 \vee ... \vee \phi'_m) \in w, \neg\langle a\rangle\phi'_1 \wedge ... \wedge \neg\langle a\rangle\phi'_m \in w & \text{(MONA)} \\
\Rightarrow & \langle a\rangle(\phi'_1 \vee ... \vee \phi'_m) \in w, \neg\langle a\rangle(\phi'_1 \vee ... \vee \phi'_m) \in w & \text{(Meta reasoning)} \\
\Rightarrow & \textbf{False} &
\end{array}
$$

Therefore, w_0 is consistent. Let w" is a maximal extension of w_0. For any ρ,

$\rho \in w' \Rightarrow \langle b\rangle\rho \in w_b \Rightarrow \langle b\rangle\rho \in w_0 \Rightarrow \langle b\rangle\rho \in w$".

So, $\langle w", w'\rangle \in R^b$.
Moreover, for any $\rho \in w$", we claim that $\langle a\rangle\rho \in w$, since if it does not hold, then we have,

$$
\begin{array}{lll}
 & \langle a\rangle\rho \notin w & \\
\Rightarrow & \neg\langle a\rangle\rho \in w & \text{(Maximal consistency of } w\text{)} \\
\Rightarrow & \neg\rho \in w_a & \text{(Definition of } w_a\text{)} \\
\Rightarrow & \neg\rho \in w_0 & \text{(Definition of } w_0\text{)} \\
\Rightarrow & \neg\rho \in w" & \text{(Definition of } w"\text{)} \\
\Rightarrow & \textbf{False} & (\rho \in w", \text{and maximal consistency of } w")
\end{array}
$$

So, $\langle w, w"\rangle \in R^a$.
Therefore, we conclude that there exists a w" such that $\langle w, w"\rangle \in R^a$ and $\langle w", w'\rangle \in R^b$. Namely, $\langle w, w'\rangle \in R^{a;b}$, which completes the proof for the case 2.

Case 3. $a' = a \cup b$,
($\Rightarrow$)

$$
\begin{array}{lll}
 & \langle w, w'\rangle \in R^{a\cup b} & \\
\Leftrightarrow & \langle w, w'\rangle \in R^a \cup R^b & \text{(Definition of } R\text{)} \\
\Leftrightarrow & \langle w, w'\rangle \in R^a or \langle w, w'\rangle \in R^b & \text{(Definition of } \cup\text{)} \\
\Leftrightarrow & \forall\rho(\rho \in w' \Rightarrow \langle a\rangle\rho \in w) \; or \; \forall\rho(\rho \in w' \Rightarrow \langle b\rangle\rho \in w) & \text{(Induction hypothesis)} \\
\Rightarrow & \forall\rho(\rho \in w' \Rightarrow \langle a\rangle\rho \in w \; or \; \langle b\rangle\rho \in w) & \text{(Meta reasoning)} \\
\Leftrightarrow & \forall\rho(\rho \in w' \Rightarrow (\langle a\rangle\rho \vee \langle b\rangle\rho) \in w) & \text{(Maximal consistency of } w\text{)} \\
\Leftrightarrow & \forall\rho(\rho \in w' \Rightarrow \langle a \cup b\rangle\rho \in w) & \text{(A4)}
\end{array}
$$

($\Leftarrow$)

$$
\begin{array}{lll}
 & \forall\rho(\rho \in w' \Rightarrow \langle a \cup b\rangle\rho \in w) & \\
\Leftrightarrow & \forall\rho(\rho \in w' \Rightarrow \langle a\rangle\rho \vee \langle b\rangle\rho \in w) & \text{(A4)}
\end{array}
$$

Case (3.a) $\langle w, w'\rangle \in R^a$
$\Rightarrow \langle w, w'\rangle \in R^{a\cup b}$.

Case (3.b) $\langle w, w'\rangle \in R^b$

$\Rightarrow \langle w, w' \rangle \in R^{a \cup b}$.

Case (3.c) $\langle w, w' \rangle \notin R^a$ and $\langle w, w' \rangle \notin R^b$
$\Rightarrow \exists \rho_1 (\rho_1 \in w'$ and $\langle a \rangle \rho_1 \notin w)$ and $\exists \rho_2 (\rho_2 \in w'$ and $\langle b \rangle \rho_2 \notin w)$ (Induction hypothesis)
$\Rightarrow \neg \langle a \rangle \rho_1 \wedge \neg \langle b \rangle \rho_2 \in w$ and $\rho_1 \wedge \rho_2 \in w'$ (Maximal consistency of w and w')
$\Rightarrow \neg (\langle a \rangle (\rho_1 \wedge \rho_2) \vee \langle b \rangle (\rho_1 \wedge \rho_2)) \in w$ and $\rho_1 \wedge \rho_2 \in w'$ (Maximal consistency of w)
$\Rightarrow \neg (\langle a \rangle (\rho_1 \wedge \rho_2) \vee \langle b \rangle (\rho_1 \wedge \rho_2)) \in w$ and $\langle a \rangle (\rho_1 \wedge \rho_2) \vee \langle b \rangle (\rho_1 \wedge \rho_2) \in w$ (Known fact)
$\Rightarrow$ **False** (Maximal consistency of w)

Therefore, we conclude that $\langle w, w' \rangle \in R^{a \cup b}$. □

Let $\phi(x)$ be a formula in which x freely occurs. In the following, we use $\phi(c)$ to denote a formula which is obtained from ϕ by replacing all free occurrences of x by c.

FACT 2 for any $w \in W$, and any valuation v, $M_c, w, v(d/x) \Vdash \phi(x) \Leftrightarrow M_c, w, v \Vdash \phi(d)$.
PROOF. First of all, we note that in the canonical model M_c, every element d

in the domain D is an interpretation of a constant $c \in CON$.
We prove the fact 2 by induction on the complexity of $\phi(x)$.

Let $v' = v(d/x)$.

Case 1. $\phi(x) = p(t_1, ..., t_{n-1}, x)$, where $p \in PRE_n$.

	$M_c, w, v(d/x) \Vdash p(t_1, ..., t_{n-1}, x)$	
$\Leftrightarrow$	$\langle v'_I(t_1), ..., v'_I(t_{n-1}), v'_I(x) \rangle \in I(p, w)$	(Truth condition)
$\Leftrightarrow$	$\langle v'_I(t_1), ..., v'_I(t_{n-1}), d \rangle \in I(p, w)$	$(v'_I(x) = d)$
$\Leftrightarrow$	$\langle v_I(t_1), ..., v_I(t_{n-1}), d \rangle \in I(p, w)$	(Definitions of v and v')
$\Leftrightarrow$	$\langle v_I(t_1), ..., v_I(t_{n-1}), v_I(d) \rangle \in I(p, w)$	$(d \in CON)$
$\Leftrightarrow$	$M_c, w, v \Vdash p(t_1, ..., t_{n-1}, d)$	(Truth condition)

Case 2. $\phi(x) = \neg \psi(x)$, and Case 3. $\phi(x) = \psi(x) \wedge \psi'(x)$.
The proofs are straightforward from the truth condition and induction hypothesis.

Case 4. $\phi(x) = \exists y \psi(y, x)$. (Note that $y \neq x$.)

	$M_c, w, v(d/x) \Vdash \exists y \psi(y, x)$	
$\Leftrightarrow$	There exists a d' such that $M_c, w, v(d/x)(d'/y) \Vdash \psi(y, x)$	(Truth condition)
$\Leftrightarrow$	There exists a d' such that $M_c, w, v(d'/y)(d/x) \Vdash \psi(y, x)$	(Definition of v)
$\Leftrightarrow$	$M_c, w, v(d'/y) \Vdash \psi(y, d)$	(Induction hypothesis)
$\Leftrightarrow$	$M_c, w, v \Vdash \exists y \phi(y, d)$	(Truth condition)

Case 5. $\phi(x) = \langle ac\rangle\psi(x)$.

$$
\begin{array}{lll}
 & M_c, w, v(d/x)\|\!-\langle ac\rangle\psi(x) & \\
\Leftrightarrow & \exists w'(R^{ac}ww' \text{ and } M_c, w, v(d/x)\|\!-\psi(x)) & \text{(Truth condition)} \\
\Leftrightarrow & \exists w'(R^{ac}ww' \text{ and } M_c, w, v\|\!-\psi(d)) & \text{(Induction hypothesis)} \\
\Leftrightarrow & M_c, w, v\|\!-\langle ac\rangle\psi(d) & \text{(Truth condition)}
\end{array}
$$

Case 6. $\phi(x) = \psi(x) \rightsquigarrow \psi'(x)$.

$$
\begin{array}{lll}
 & M_c, w, v(d/x)\|\!-\psi(x) \rightsquigarrow \psi'(x) & \\
\Leftrightarrow & cw(w, [\![\psi(x)]\!]^{v(d/x)}_{M_c}) \subseteq [\![\psi'(x)]\!]^{v(d/x)}_{M_c} & \text{(Truth condition)} \\
\Leftrightarrow & cw(w, [\![\psi(d)]\!]^{v}_{M_c}) \subseteq [\![\psi'(d)]\!]^{v}_{M_c} & \text{(Induction hypothesis)} \\
\Leftrightarrow & M_c, w, v\|\!-\psi(d) \rightsquigarrow \psi'(d) & \text{(Truth condition)}
\end{array}
$$

Case 7. $\phi(x) = \psi(x)\boldsymbol{P}_{ag}\psi'(x)$.

$$
\begin{array}{lll}
 & M_c, w, v(d/x)\|\!-\psi(x)\boldsymbol{P}_{ag}\psi'(x) & \\
\Leftrightarrow & cw(w, [\![\psi(x) \wedge \neg\psi'(x)]\!]^{v(d/x)}_{M_c}) \succ_{ag} cw(w, [\![\psi'(x) \wedge \neg\psi(x)]\!]^{v(d/x)}_{M_c}) & \text{(Truth condition)} \\
\Leftrightarrow & cw(w, [\![\psi(d) \wedge \neg\psi'(d)]\!]^{v}_{M_c}) \succ_{ag} cw(w, [\![\psi'(d) \wedge \neg\psi(d)]\!]^{v}_{M_c}) & \text{(Induction hypothesis)} \\
\Leftrightarrow & M_c, w, v\|\!-\psi(d)\boldsymbol{P}_{ag}\psi'(d) & \text{(Truth condition)}
\end{array}
$$

Case 8. $\phi(x) = \mathbf{B}_{ag}\psi(x)$.

$$
\begin{array}{lll}
 & M_c, w, v(d/x)\|\!-\mathbf{B}_{ag}\psi(x) & \\
\Leftrightarrow & \forall w'(\mathcal{B}_{ag}ww' \Rightarrow M_c, w', v(d/x)\|\!-\psi(x)) & \text{(Truth condition)} \\
\Leftrightarrow & \forall w'(\mathcal{B}_{ag}ww' \Rightarrow M_c, w', v\|\!-\psi(d)) & \text{(Induction hypothesis)} \\
\Leftrightarrow & M_c, w, v\|\!-\mathbf{B}_{ag}\psi(d) &
\end{array}
$$

□

FACT 3(Witness property): For the canonical model M_c, the valuation v, a formula $\phi(x)$, and any world $w \in W$, it holds that
$\exists x\phi \in w \Rightarrow \phi(c) \in w$ for some constant $c \in CON$.

PROOF. First, we claim that,

for any $w \in W$, any valuation v, $M_c, w, v\|\!-\exists x\phi(x) \rightarrow \phi(c)$,
for some constant $c \in CON$.

$$
\begin{array}{lll}
 & M_c, w, v \Vdash \exists x \phi(x) & \\
\Leftrightarrow & M_c, w, v(d/x) \Vdash \phi(x), \text{for some } d \in D, & \text{(Truth condition)} \\
\Leftrightarrow & M_c, w, v(d/x) \Vdash \phi(x), \text{for some } d \in CON, & \text{(Domain } D = CON \text{ in } M_c) \\
\Leftrightarrow & M_c, w, v \Vdash \phi(d), \text{for some } d \in CON, & \text{(FACT 2)}
\end{array}
$$

Therefore, we know that $\exists x \phi(x) \rightarrow \phi(c)$, for some $c \in CON$, is valid with respect to the canonical model M_c. Therefore, $\exists x \phi(x) \in w \Rightarrow \phi(c) \in w$, for some $c \in CON$. □

Now, we are ready to prove the truth lemma by induction on the complexity of χ.

(1) $\chi \in w \Leftrightarrow w \in [\![\chi]\!]^v_{M_c}$.

(1.1) $\chi \equiv p(t_1, ..., t_n)$ where $p(t_1, ..., t_n) \in ATOM, p \in PRE_n$,

$$
\begin{array}{lll}
 & p(t_1, ..., t_n) \in w & \\
\Leftrightarrow & \langle v_I(t_1), ..., v_I(t_n) \rangle \in I_P(p, w) & \text{(Definition of } I) \\
\Leftrightarrow & w \in [\![p(t_1, ..., t_n)]\!]^v_{M_c} & \text{(Truth condition)}
\end{array}
$$

(1.2) $\chi \equiv \neg \phi$.

$$
\begin{array}{lll}
 & \neg \phi \in w & \\
\Leftrightarrow & \phi \notin w & \text{(Maximal consistency of } w) \\
\Leftrightarrow & w \notin [\![\phi]\!]^v_{M_c} & \text{(Induction hypothesis)} \\
\Leftrightarrow & w \in [\![\neg \phi]\!]^v_{M_c} & \text{(Truth condition)}
\end{array}
$$

(1.3) $\chi \equiv \phi \wedge \psi$.

$$
\begin{array}{lll}
 & \phi \wedge \psi \in w & \\
\Leftrightarrow & \phi, \psi \in w & \text{(Maximal consistency of } w) \\
\Leftrightarrow & w \in [\![\phi]\!]^v_{M_c} \text{ and } w \in [\![\psi]\!]^v_{M_c} & \text{(Induction hypothesis)} \\
\Leftrightarrow & w \in [\![\phi \wedge \psi]\!]^v_{M_c} & \text{(Truth condition)}
\end{array}
$$

(1.4) $\chi \equiv \exists x \phi(x)$.
(⇒)

$$
\begin{array}{lll}
 & \exists x \phi(x) \in w & \\
\Rightarrow & \phi(c) \in w & \text{(Witness Property)} \\
\Leftrightarrow & M_c, w, v \Vdash \phi(c) & \text{(Induction hypothesis)} \\
\Rightarrow & M_c, w, v \Vdash \exists x \phi(x) & \text{(First order logic)}
\end{array}
$$

($\Leftarrow$)

$$\begin{array}{lll} & M_c, w, v \| \!\!- \exists x \phi(x) & \\ \Leftrightarrow & M_c, w, v(d/x) \| \!\!- \phi(x) & \text{(Truth condition)} \\ \Rightarrow & M_c, w, v \| \!\!- \phi(d) & \text{(FACT 2)} \\ \Leftrightarrow & \phi(d) \in w & \text{(Induction hypothesis)} \\ \Rightarrow & \exists x \phi \in w & \text{(First order logic)} \end{array}$$

(1.5) $\chi \equiv \langle ac \rangle \phi$.
($\Rightarrow$)

$$\begin{array}{lll} & \langle ac \rangle \phi \in w & \\ \Rightarrow & (\exists w' \in W)(\phi \in w' \text{ and } \forall \psi \in w'(\langle ac \rangle \psi \in w)) & \text{(Action lemma)} \\ \Rightarrow & \exists w'(\phi \in w' \text{ and} \langle w, w' \rangle \in R^{ac}) & \text{(FACT 1)} \\ \Rightarrow & \exists w'(w' \in [\![\phi]\!]^v_{M_c} \text{ and } \langle w, w' \rangle \in R^{ac}) & \text{(Induction hypothesis)} \\ \Rightarrow & w \in [\![\langle ac \rangle \phi]\!]^v_{M_c} & \text{(Truth condition)} \end{array}$$

($\Leftarrow$)

$$\begin{array}{lll} & w \in [\![\langle ac \rangle \phi]\!]^v_{M_c} & \\ \Leftrightarrow & (\exists w' \in W)(\langle w, w' \rangle \in R^{ac}) \text{ and } w' \in [\![\phi]\!]^v_{M_c}) & \text{(Truth condition)} \\ \Leftrightarrow & (\exists w' \in W)(\langle w, w' \rangle \in R^{ac} \text{ and } \phi \in w') & \text{(Induction hypothesis)} \\ \Rightarrow & \langle ac \rangle \phi \in w & \text{(FACT 1)} \end{array}$$

(1.6) $\chi \equiv \phi \rightsquigarrow \psi$.

$$\begin{array}{lll} & \phi \rightsquigarrow \psi \in w & \\ \Rightarrow & \forall w'(w' \in cw(w, [\![\phi]\!]^v_{M_c}) \Rightarrow \psi \in w') & \text{(Definition of } cw) \\ \Rightarrow & \forall w'(w' \in cw(w, [\![\phi]\!]^v_{M_c}) \Rightarrow M_c, w', v \| \!\!- \psi) & \text{(Induction hypothesis)} \\ \Leftrightarrow & w \in [\![\phi \rightsquigarrow \psi]\!]^v_{M_c} & \text{(Truth condition)} \end{array}$$

($\Leftarrow$)

$$\begin{array}{lll} & \neg(\phi \rightsquigarrow \psi) \in w & \\ \Rightarrow & \exists w'(\neg\psi \in w' \text{ and } \forall \rho((\phi \rightsquigarrow \rho) \in w \Rightarrow \rho \in w')) & \text{(conditional lemma)} \\ \Leftrightarrow & \exists w'(\neg\psi \in w' \text{ and } w' \in cw(w, [\![\phi]\!]^v_{M_c}) & \text{(Definition of } cw) \\ \Leftrightarrow & M_c, w, v \| \!\!\not- \phi \rightsquigarrow \psi & \text{(Truth condition)} \end{array}$$

(1.7) $\chi \equiv \phi \boldsymbol{P}_{ag} \psi$.

$$\begin{array}{lll} & \phi \boldsymbol{P}_{ag} \psi \in w & \\ \Leftrightarrow & cw(w, [\![\phi \wedge \neg\psi]\!]^v_{M_c}) \succ_{v_I(ag)} cw(w, [\![\psi \wedge \neg\phi]\!]^v_{M_c}) & \text{(Definition of } \succ) \\ \Leftrightarrow & w \in [\![\phi \boldsymbol{P}_{ag} \psi]\!]^v_{M_c} & \text{(Truth condition)} \end{array}$$

(1.8) $\chi \equiv \boldsymbol{B}_{ag} \phi$.

$$
\begin{array}{lll}
 & \boldsymbol{B}_{ag}\phi \in w & \\
\Rightarrow & \forall w'(\langle w, w'\rangle \in \mathcal{B}(v_I(ag)) \Rightarrow \phi \in w') & \text{(Definition of } \mathcal{B}) \\
\Rightarrow & \forall w'(\langle w, w'\rangle \in \mathcal{B}(v_I(ag)) \Rightarrow M_c, w', v \Vdash \phi) & \text{(Induction hypothesis)} \\
\Leftrightarrow & w \in [\![\boldsymbol{B}_{ag}\phi]\!]^v_{M_c} & \text{(Truth condition)}
\end{array}
$$

($\Leftarrow$)

$$
\begin{array}{lll}
 & \neg(\boldsymbol{B}_{ag}\phi) \in w & \\
\Rightarrow & \exists w'(\neg\phi \in w' \text{ and } \forall\rho((\boldsymbol{B}_{ag}\rho) \in w \Rightarrow \rho \in w')) & \text{(Belief lemma)} \\
\Leftrightarrow & \exists w'(\neg\phi \in w' \text{ and } \langle w, w'\rangle \in \mathcal{B}(v_I(ag))) & \text{(Definition of } \mathcal{B}) \\
\Leftrightarrow & M_c, w, v \nVdash \boldsymbol{B}_{ag}\phi & \text{(Truth condition)}
\end{array}
$$

This concludes the proof of the truth lemma. We now show that M_c is an ALX3 model. So, we have to show that cw satisfies (CS1)-(CS5), $\succ$ satisfies the normality and transitivity conditions, and $\mathcal{B}$ is serial and transitive.

(CS1) $w \in cw(w', [\![\psi]\!]^v_{M_c}) \Rightarrow w \in [\![\psi]\!]^v_{M_c}$.

$$
\begin{array}{lll}
 & w \in cw(w', [\![\psi]\!]_{M_c}) & \\
\Leftrightarrow & \forall\rho((\psi \rightsquigarrow \rho) \in w' \Rightarrow \rho \in w) & \text{(Definition of } cw) \\
\Rightarrow & \psi \in w & \text{(ID)} \\
\Leftrightarrow & w \in [\![\psi]\!]^v_{M_c} & \text{(Truth lemma)}
\end{array}
$$

(CS2) $w \in [\![\psi]\!]^v_{M_c} \Rightarrow cw(w, [\![\psi]\!]^v_{M_c}) = \{w\}$.

we must show that:
(a) $w \in [\![\psi]\!]^v_{M_c} \Rightarrow w \in cw(w, [\![\psi]\!]^v_{M_c})$,
(b) $w \in [\![\psi]\!]^v_{M_c}$ and $w' \in cw(w, [\![\psi]\!]^v_{M_c}) \Rightarrow w = w'$.

For (a), we have:

$$
\begin{array}{lll}
 & w \in [\![\psi]\!]^v_{M_c} & \\
\Leftrightarrow & \psi \in w & \text{(Truth lemma)} \\
\Rightarrow & \forall\rho(\psi \rightsquigarrow \rho \in w \Rightarrow \rho \in w) & \text{(MPC)} \\
\Rightarrow & w \in cw(w, [\![\psi]\!]^v_{M_c}) & \text{(Definition of } cw)
\end{array}
$$

For (b), suppose that $w \in [\![\psi]\!]^v_{M_c}$ and $w' \in cw(w, [\![\psi]\!]^v_{M_c})$, we first show that $w \subseteq w'$. Then by the maximal consistency of both w and w', we have $w = w'$. To show that $w \subseteq w'$, we proceed by reductio ad absurdum and show that $\rho \in w$ and $\rho \notin w' \Rightarrow$ **False**, for arbitrary ρ.

	$\rho \in w$ and $\rho \notin w'$	
$\Leftrightarrow$	$\rho \in w$ and $\neg\rho \in w'$	(Maximal consistency of w')
$\Rightarrow$	$\rho \wedge \psi \in w$ and $\neg\rho \in w'$	($w \in [\![\psi]\!]^v_{M_c}$ and maximal consistency of w)
$\Rightarrow$	$(\psi \rightsquigarrow \rho) \in w$ and $\neg\rho \in w'$	(CS)
$\Rightarrow$	$\rho \in w'$ and $\neg\rho \in w'$	($w' \in cw(j, [\![\psi]\!]^v_{M_c})$ and definition of cw)
$\Rightarrow$	**False**	(Maximal consistency of w')

(CS3) $cw(w, [\![\psi]\!]^v_{M_c}) = \emptyset \Rightarrow cw(w, [\![\phi]\!]^v_{M_c}) \cap [\![\psi]\!]^v_{M_c} = \emptyset$.

We show the contraposition of (CS3), namely,

$cw(w, [\![\phi]\!]^v_{M_c}) \cap [\![\psi]\!]^v_{M_c} \neq \emptyset \Rightarrow cw(w, [\![\psi]\!]^v_{M_c}) \neq \emptyset$.

	$cw(w, [\![\phi]\!]^v_{M_c}) \cap [\![\psi]\!]^v_{M_c} \neq \emptyset$	
$\Rightarrow$	$\exists w'(w' \in cw(w, [\![\phi]\!]^v_{M_c}) \cap [\![\psi]\!]^v_{M_c})$	(Meta reasoning)
$\Rightarrow$	$w' \in cw(w, [\![\phi]\!]^v_{M_c})$ and $w' \in [\![\psi]\!]^v_{M_c}$	(Meta reasoning)
$\Rightarrow$	$w' \in cw(w, [\![\phi]\!]^v_{M_c})$ and $\psi \in w'$	(Truth lemma)
$\Rightarrow$	$(\phi \rightsquigarrow \neg\psi) \notin w$	(Maximal consistency of w' and definition of cw)
$\Rightarrow$	$\neg(\phi \rightsquigarrow \neg\psi) \in w$	(Maximal consistency of w)
$\Rightarrow$	$\neg(\psi \rightsquigarrow \neg\psi) \in w$	(MOD)

Let $w^\circ = \{\rho : \psi \rightsquigarrow \rho \in w\}$.

We show that w° is consistent.

	$\bot \in w^\circ$	
$\Rightarrow$	$\exists\rho_1 ... \rho_n((\psi \rightsquigarrow \rho_1) \in w$ and ... and $(\psi \rightsquigarrow \rho_n)$ and $\vdash \rho_1 \wedge ... \wedge \rho_n \rightarrow \bot)$	(Definition of w°)
$\Rightarrow$	$(\psi \rightsquigarrow \rho_1 \wedge ... \wedge \rho_n) \in w$ and $\vdash (\rho_1 \wedge ... \wedge \rho_n \rightarrow \bot)$	(CC)
$\Rightarrow$	$\psi \rightsquigarrow \bot \in w$	(Maximal consistency of w)
$\Rightarrow$	$(\psi \rightsquigarrow \neg\psi) \in w$	(MONC)
$\Rightarrow$	**False**	(Maximal consistency of w)

Therefore, w° is consistent. Let w' be one of maximal extensions of w°
Thus, we have $w' \in cw(w, [\![\psi]\!]^v_{M_c})$ from the construction of w and the definition of cw. So (CS3) holds.

(CS4) $cw(w, [\![\phi]\!]^v_{M_c}) \subseteq [\![\psi]\!]^v_{M_c}$ and $cw(w, [\![\psi]\!]^v_{M_c}) \subseteq [\![\phi]\!]^v_{M_c} \Rightarrow cw(w, [\![\phi]\!]^v_{M_c}) = cw(w, [\![\psi]\!]^v_{M_c})$.

Assume that $cw(w, [\![\phi]\!]^v_{M_c}) \subseteq [\![\psi]\!]^v_{M_c}$ and $cw(w, [\![\psi]\!]^v_{M_c}) \subseteq [\![\phi]\!]^v_{M_c}$. Then we have $M_c, w, v \Vdash (\phi \rightsquigarrow \psi) \wedge (\psi \rightsquigarrow \phi)$ by the truth condition. Furthermore, by the truth lemma, we have $((\phi \rightsquigarrow \psi) \wedge (\psi \rightsquigarrow \phi)) \in w$. For any $w' \in cw(w, [\![\phi]\!]^v_{M_c})$, we want to prove that $w' \in cw(w, [\![\psi]\!]^v_{M_c})$.

For any formula ρ,

$$\begin{array}{lll} & (\psi \rightsquigarrow \rho) \in w & \\ \Rightarrow & ((\phi \rightsquigarrow \psi) \wedge (\psi \rightsquigarrow \phi) \wedge (\psi \rightsquigarrow \rho)) \in w & ((\phi \rightsquigarrow \psi) \wedge (\psi \rightsquigarrow \phi) \in w) \\ \Rightarrow & (\phi \rightsquigarrow \rho) \in w & (\text{CSO}) \\ \Rightarrow & \rho \in w' & (w' \in cw(w, [\![\phi]\!]^v_{M_c}) \text{ and definition of } cw) \end{array}$$

Therefore, by the definition of cw, we have $w' \in cw(w, [\![\psi]\!]^v_{M_c})$. So, $cw(w, [\![\phi]\!]^v_{M_c}) \subseteq cw(w, [\![\psi]\!]^v_M)$. Similarly, we can prove that $cw(w, [\![\psi]\!]^v_{M_c}) \subseteq cw(w, [\![\phi]\!]^v_{M_c})$, which concludes (CS4).

(CS5) $cw(w, [\![\phi]\!]^v_{M_c}) \cap [\![\psi]\!]^v_{M_c} \neq \emptyset \Rightarrow cw(w, [\![\psi \wedge \phi]\!]^v_{M_c}) \subseteq cw(w, [\![\phi]\!]^v_{M_c})$.

Assume that $cw(w, [\![\phi]\!]^v_{M_c}) \cap [\![\psi]\!]^v_{M_c} \neq \emptyset$. Then by the truth condition, we have $M_c, w, v \Vdash \neg(\phi \rightsquigarrow \neg\psi)$. Furthermore, by the truth lemma, we have $\neg(\phi \rightsquigarrow \neg\psi) \in w$. Now, we have to show that for any $w' \in cw(w, [\![\phi \wedge \psi]\!]^v_{M_c})$, $w' \in cw(w, [\![\phi]\!]^v_{M_c})$

For any formula ρ,

$$\begin{array}{lll} & \phi \rightsquigarrow \rho \in w & \\ \Rightarrow & ((\phi \rightsquigarrow \rho) \wedge \neg(\phi \rightsquigarrow \neg\psi)) \in w & (\neg(\phi \rightsquigarrow \neg\psi) \in w) \\ \Rightarrow & ((\phi \wedge \psi) \rightsquigarrow \rho) \in w & (\text{CV}) \\ \Rightarrow & \rho \in w' & (w' \in cw(w, [\![\phi \wedge \psi]\!]^v_{M_c}) \text{ and definition of } cw) \end{array}$$

Therefore, $w' \in cw(w, [\![\phi]\!]^v_{M_c})$ by the definition of cw, so (CS5) holds.

Note that (CEP), (N) and (TR) are axioms of ALX3. Similar to the proofs for the completeness of ALX1, it is easy to prove that (NORM) and (TRAN) also hold for the canonical model of ALX3. Here we do not want to go into the details.

Finally, we have to prove that $\mathcal{B}(ag)$ is serial and transitive, namely, (SEB) and (TRB) hold in the canonical model M_c.

(SEB) $\forall w \exists w'(\langle w, w' \rangle \in \mathcal{B}(ag))$.

For any $w \in W$, and any $ag \in AGENT$,
let $w^* = \{\rho : \mathbf{B}_{ag}\rho \in w\}$.
Since (DB) is valid in ALX3S, it is easy to see that w^* is consistent. Let w' is the maximal extension of w^*. Then, it is easy to see that w' is a world which we need. Therefore, (SEB) holds in M_c.

(TRB) $\mathcal{B}(ag) \circ \mathcal{B}(ag) \subseteq \mathcal{B}(ag)$.

Since (TR) is valid in ALX3S, it is easy to see that (TRB) hold in M_c. Here we do not want to go into details. See [Hughes 1984] for the details.

This concludes the proof that M_c is an ALX3 model. □

11.5 More Operations

Based on the primitive system of ALX3, we can define more modal operations, as illustrated before. Defining more operations makes ALX3 logic more expressive and more flexible. One of the strengths of ALX3 is to introduce the operations of necessity and possibility freely from the conditional.

11.5.1 Necessity and Possibility

We define the operations of necessity and possibility in terms of the conditional as follows:

$$\Box\phi \overset{\text{def}}{\Longleftrightarrow} \neg\phi \rightsquigarrow \phi.$$

$$\Diamond\phi \overset{\text{def}}{\Longleftrightarrow} \neg(\phi \rightsquigarrow \neg\phi).$$

11.5.1. Lemma. *For any model* $M = \langle O, PA, AGENT, W, cw, \succ, \mathcal{R}, \mathcal{B}, I\rangle$, *any* $w, w' \in W$, *any* $\phi, \psi \in FML$, *and any variable valuation* $v \in V_D$,

(i) $M, w, v\Vdash\!\!-\Box\phi \Leftrightarrow cw(w, [\![\neg\phi]\!]^v_M) = \emptyset$.
(ii) $w' \in cw(w, [\![\psi]\!]^v_M)$ *and* $cw(w, [\![\phi]\!]^v_M) = \emptyset \Rightarrow w' \notin [\![\phi]\!]^v_M$.

PROOF.
(i) (⇒)

$$\begin{array}{lll}
 & M, w, v\Vdash\!\!-\Box\phi & \\
\Leftrightarrow & M, w, v\Vdash\!\!-\neg\phi \rightsquigarrow \phi & \text{(Definition of } \Box) \\
\Leftrightarrow & cw(w, [\![\neg\phi]\!]^v_M) \subseteq [\![\phi]\!]^v_M & \text{(Truth condition)} \\
\Rightarrow & cw(w, [\![\neg\phi]\!]^v_M) \subseteq [\![\neg\phi]\!]^v_M \cap [\![\phi]\!]^v_M & \text{(CS1)} \\
\Rightarrow & cw(w, [\![\neg\phi]\!]^v_M) \subseteq \emptyset & \text{(Meta reasoning)} \\
\Rightarrow & cw(w, [\![\neg\phi]\!]^v_M) = \emptyset & \text{(Meta reasoning)}
\end{array}$$

(⇐)

$$\begin{array}{lll}
 & cw(w, [\![\neg\phi]\!]^v_M) = \emptyset & \\
\Rightarrow & cw(w, [\![\neg\phi]\!]^v_M) \subseteq [\![\phi]\!]^v_M & \text{(Meta reasoning)} \\
\Leftrightarrow & M, w, v\Vdash\!\!-\neg\phi \rightsquigarrow \phi & \text{(Truth condition)} \\
\Leftrightarrow & M, w, v\Vdash\!\!-\Box\phi & \text{(Definition of } \Box)
\end{array}$$

(ii) Suppose that $w' \in cw(w, [\![\psi]\!]^v_M)$ and $cw(w, [\![\phi]\!]^v_M) = \emptyset$.

We have to show that $w' \in [\![\phi]\!]^v_M \Rightarrow$ **False**.

$$
\begin{array}{lll}
 & w' \in cw(w[\![\psi]\!]^v_M) \text{ and } cw(w, [\![\phi]\!]^v_M) = \emptyset \text{ and } w' \in [\![\phi]\!]^v_M & \\
\Rightarrow & w' \in cw(w, [\![\psi]\!]^v_M) \cap [\![\phi]\!]^v_M \text{ and } cw(w, [\![\phi]\!]^v_M) = \emptyset & \text{(Meta reasoning)} \\
\Rightarrow & cw(w, [\![\psi]\!]^v_M) \cap [\![\phi]\!]^v_M \neq \emptyset \text{ and } cw(w, [\![\phi]\!]^v_M) = \emptyset & \text{(Meta reasoning)} \\
\Rightarrow & cw(w, [\![\phi]\!]^v_M) \neq \emptyset \text{ and } cw(w, [\![\phi]\!]^v_M) = \emptyset & \text{(CS3)} \\
\Rightarrow & \textbf{False} & \text{(Meta reasoning)}
\end{array}
$$

□

In the following propositions, we show that the above definition gives us axioms (K) and (T) for the necessity operation. Moreover, if we need more axioms such as (4) and (5), we need additional conditions on the closest world function.

11.5.1. Proposition. *The following axioms about the necessity operator and possibility operator are sound for the class of ALX3 models.*

(K) $\Box(\phi \rightarrow \psi) \wedge \Box\phi \rightarrow \Box\psi$.
(T) $\Box\phi \rightarrow \phi$.
(D) $\Box\phi \rightarrow \Diamond\phi$.
(SC) $\Box(\phi \rightarrow \psi) \rightarrow (\phi \rightsquigarrow \psi)$.
(NP) $\Diamond\phi \leftrightarrow \neg\Box\neg\phi$.
(BF) $\forall x \Box\phi \rightarrow \Box\forall x\phi$.

PROOF. Let $M = \langle O, PA, AGENT, W, cw, \succ, \mathcal{R}, \mathcal{B}, I\rangle$.
For any $w, w' \in W$, $\phi, \psi \in FML$, and variable valuation $v \in V_D$,

(K) $\Box(\phi \rightarrow \psi) \wedge \Box\phi \rightarrow \Box\psi$.

Suppose that $M, w, v\|\!\!-\Box(\phi \rightarrow \psi) \wedge \Box\phi$. For any $w' \in cw(w, [\![\neg\psi]\!]^v_M)$,

$$
\begin{array}{lll}
 & M, w, v\|\!\!-\Box(\phi \rightarrow \psi) \wedge \Box\phi \text{ and } w' \in cw(w, [\![\neg\psi]\!]^v_M) & \\
\Rightarrow & cw(w, [\![\neg(\phi \rightarrow \psi)]\!]^v_M) = \emptyset \text{ and } cw(w, [\![\neg\phi]\!]^v_M) = \emptyset & \\
 & \text{and } w' \in cw(w, [\![\neg\psi]\!]^v_M) & \text{(Lemma (i))} \\
\Rightarrow & w' \notin [\![\neg(\phi \rightarrow \psi)]\!]^v_M \text{ and } w' \notin [\![\neg\phi]\!]^v_M & \text{(Lemma (ii))} \\
\Rightarrow & w' \in [\![(\phi \rightarrow \psi)]\!]^v_M \text{ and } w' \in [\![\phi]\!]^v_M & \text{(Meta reasoning)} \\
\Rightarrow & w' \in [\![\psi]\!]^v_M & \text{(Meta reasoning)}
\end{array}
$$

So, $M, w, v\|\!\!-(\neg\psi \rightsquigarrow \psi)$, and consequently $M, w, v\|\!\!-\Box\psi$.

(T) $\Box\phi \rightarrow \phi$.

$$
\begin{array}{lll}
 & M, w, v \| \!\!- \Box\phi & \\
\Leftrightarrow & M, w, v \| \!\!- \neg\phi \rightsquigarrow \phi & \text{(Definition of } \Box\text{)} \\
\Rightarrow & M, w, v \| \!\!- \neg\phi \rightarrow \phi & \text{(MPC)} \\
\Leftrightarrow & M, w, v \| \!\!- \phi \vee \phi & \text{(Meta reasoning)} \\
\Leftrightarrow & M, w, v \| \!\!- \phi & \text{(Meta reasoning)}
\end{array}
$$

(D) $\Box\phi \rightarrow \Diamond\phi$.

$$
\begin{array}{lll}
 & M, w, v \| \!\!- \Box\phi & \\
\Rightarrow & M, w, v \| \!\!- \phi & \text{(T)} \\
\Rightarrow & M, w, v \| \!\!- \neg(\phi \rightarrow \neg\phi) & \text{(Meta reasoning)} \\
\Rightarrow & M, w, v \| \!\!- \neg(\phi \rightsquigarrow \neg\phi) & \text{(MPC)} \\
\Rightarrow & M, w, v \| \!\!- \Diamond\phi & \text{(Definition of } \Diamond\text{)}
\end{array}
$$

(SC) $\Box(\phi \rightarrow \psi) \rightarrow (\phi \rightsquigarrow \psi)$.

Suppose that $M, w, v \| \!\!- \Box(\phi \rightarrow \psi)$. For any $w' \in cw(w, [\![\phi]\!]^v_M)$,

$$
\begin{array}{lll}
 & M, w, v \| \!\!- \Box(\phi \rightarrow \psi) \text{ and } w' \in cw(w, [\![\phi]\!]^v_M) & \\
\Rightarrow & cw(w, [\![\neg(\phi \rightarrow \psi)]\!]^v_M) = \emptyset \text{ and } w' \in cw(w, [\![\phi]\!]^v_M) & \text{(Lemma (i))} \\
\Rightarrow & cw(w, [\![\neg(\phi \rightarrow \psi)]\!]^v_M) = \emptyset \text{ and } w' \in [\![\phi]\!]^v_M & \text{(CS1)} \\
\Rightarrow & w' \notin [\![\neg(\phi \rightarrow \psi)]\!]^v_M \text{ and } w' \in [\![\phi]\!]^v_M & \text{(Lemma (ii))} \\
\Rightarrow & w' \in [\![(\phi \rightarrow \psi)]\!]^v_M \text{ and } w' \in [\![\phi]\!]^v_M & \text{(Meta reasoning)} \\
\Rightarrow & w' \in [\![\psi]\!]^v_M & \text{(Meta reasoning)}
\end{array}
$$

So, $M, w, v \| \!\!- (\phi \rightsquigarrow \psi)$.

(NP) $\Diamond\phi \leftrightarrow \neg\Box\neg\phi$.

$$
\begin{array}{lll}
 & M, w, v \| \!\!- \Diamond\phi & \\
\Leftrightarrow & M, w, v \| \!\!- \neg(\phi \rightsquigarrow \neg\phi) & \text{(Definition of } \Diamond\text{)} \\
\Leftrightarrow & M, w, v \| \!\!- \neg(\neg(\neg\phi) \rightsquigarrow \neg\phi) & \text{(Meta reasoning)} \\
\Leftrightarrow & M, w, v \| \!\!- \neg\Box\neg\phi & \text{(Definition of } \Box\text{)}
\end{array}
$$

(BF) $\forall x \Box\phi \rightarrow \Box\forall x\phi$.

	$M, w, v \Vdash \forall x \Box \phi$	
$\Leftrightarrow$	$\forall d(M, w, v(d/x) \Vdash \Box \phi)$	(Truth condition)
$\Leftrightarrow$	$\forall d(M, w, v(d/x) \Vdash \neg \phi \rightsquigarrow \phi)$	(Definition of $\Box$)
$\Leftrightarrow$	$\forall d(cw(w, [\![\neg \phi]\!]_M^{v(d/x)}) \subseteq [\![\phi]\!]_M^{v(d/x)})$	(Truth condition)
$\Rightarrow$	$\forall d(cw(w, [\![\neg \phi]\!]_M^{v(d/x)}) = \emptyset)$	(CS1 and Meta Reasoning)
$\Rightarrow$	$cw(w, [\![\neg \phi]\!]_M^{v(d/x)}) = \emptyset$	(Meta reasoning)
$\Rightarrow$	$cw(w, [\![\exists x \neg \phi]\!]_M^{v}) = \emptyset$	$([\![\neg \phi]\!]_M^{v(d/x)} = [\![\exists x \neg \phi]\!]_M^{v})$
$\Rightarrow$	$cw(w, [\![\neg \phi]\!]_M^{v}) \subseteq [\![\neg \exists x \neg \phi]\!]_M^{v}$	(Meta reasoning)
$\Leftrightarrow$	$M, w, v \Vdash \exists x \neg \phi \rightsquigarrow \neg \exists \neg \phi$	(Truth condition)
$\Leftrightarrow$	$M, w, v \Vdash \neg(\neg \exists x \neg \phi) \rightsquigarrow \neg \exists x \neg \phi$	(Meta reasoning)
$\Leftrightarrow$	$M, w, v \Vdash \Box \neg \exists x \neg \phi$	(Definition of $\Box$)
$\Leftrightarrow$	$M, w, v \Vdash \Box \forall x \phi$	(Meta reasoning)

□

11.5.2. Proposition. *For any model* $M = \langle O, PA, AGENT, W, cw, \succ, \mathcal{R}, \mathcal{B}, I \rangle$*, if the closest world function cw satisfies (CSN), then the following axioms are sound for the class of the semantics models:*
(4) $\Box \phi \rightarrow \Box \Box \phi$.
(5) $\neg \Box \phi \rightarrow \Box \neg \Box \phi$.

PROOF. (4) $\Box \phi \rightarrow \Box \Box \phi$.

	$M, w, v \Vdash \Box \phi$	
$\Rightarrow$	$cw(w, [\![\neg \phi]\!]_M^{v}) = \emptyset$	(Lemma (i))
$\Rightarrow$	$[\![\neg \phi]\!]_M^{v} = \emptyset$	(CSN)
$\Rightarrow$	$\forall w'(cw(w', [\![\neg \phi]\!]_M^{v}) = \emptyset)$	(CS1)
$\Rightarrow$	$\forall w'(cw(w', [\![\neg \phi]\!]_M^{v}) \subseteq [\![\phi]\!]_M^{v})$	(Meta reasoning)
$\Rightarrow$	$\forall w'(M, w', v \Vdash \neg \phi \rightsquigarrow \phi)$	(Truth condition)
$\Rightarrow$	$[\![\neg \phi \rightsquigarrow \phi]\!]_M^{v} = W$	(Definition of $[\![\ ]\!]_M^{v}$)
$\Rightarrow$	$[\![\Box \phi]\!]_M^{v} = W$	(Definition of $\Box$)
$\Rightarrow$	$[\![\neg \Box \phi]\!]_M^{v} = \emptyset$	(Meta reasoning)
$\Rightarrow$	$cw(w, [\![\neg \Box \phi]\!]_M^{v}) = \emptyset$	(CS1)
$\Rightarrow$	$cw(w, [\![\neg \Box \phi]\!]_M^{v}) \subseteq [\![\Box \phi]\!]_M^{v}$	(Meta reasoning)
$\Rightarrow$	$M, w, v \Vdash \neg \Box \phi \rightsquigarrow \Box \phi$	(Truth condition)
$\Rightarrow$	$M, w, v \Vdash \Box \Box \phi$	(Definition of $\Box$)

(5) $\neg \Box \phi \rightarrow \Box \neg \Box \phi$.

$$
\begin{array}{lll}
 & M, w, v \Vdash \neg\Box\phi & \\
\Rightarrow & cw(w, [\![\neg\phi]\!]^v_M) \neq \emptyset & \text{(Lemma (i))} \\
\Rightarrow & [\![\neg\phi]\!]^v_M \neq \emptyset & \text{(CS1)} \\
\Rightarrow & \forall w'(cw(w', [\![\neg\phi]\!]^v_M) \neq \emptyset) & \text{(CSN)} \\
\Rightarrow & \forall w'(cw(w', [\![\neg\phi]\!]^v_M) \not\subseteq [\![\phi]\!]^v_M & \text{(Meta reasoning)} \\
\Rightarrow & \forall w'(M, w', v \Vdash \neg\Box\phi & \text{(Truth condition)} \\
\Rightarrow & [\![\Box\phi]\!]^v_M = \emptyset & \text{(Definition of } [\![\ ]\!]^v_M) \\
\Rightarrow & cw(w, [\![\Box\phi]\!]^v_M) = \emptyset & \text{(CS1)} \\
\Rightarrow & cw(w, [\![\Box\phi]\!]^v_M) \subseteq [\![\neg\Box\phi]\!]^v_M & \text{(Meta reasoning)} \\
\Rightarrow & M, w, v \Vdash \Box\phi \rightsquigarrow \neg\Box\phi & \text{(Truth condition)} \\
\Rightarrow & M, w, v \Vdash \Box(\neg\Box\phi) & \text{(Definition of } \Box)
\end{array}
$$

□

11.5.2 Beliefs and Knowledge

11.5.3. Proposition. (More properties of the belief operator) *The following axioms for the belief operator are sound for the class of ALX3 models.*

(B∧) $\boldsymbol{B}_i\phi \wedge \boldsymbol{B}_i\psi \leftrightarrow \boldsymbol{B}_i(\phi \wedge \psi)$.
(B∨) $\boldsymbol{B}_i\phi \vee \boldsymbol{B}_i\psi \rightarrow \boldsymbol{B}_i(\phi \vee \psi)$.

PROOF. Straightforward from the meaning functions. □

It is interesting that we can define a knowledge operator in terms of the belief operation; this idea originates from [Cohen&Levesque 1987]. In the following, we examine the problem from the perspective of ALX logic.

We define the knowledge operator $\boldsymbol{K}_i$ as follows:

$$\boldsymbol{K}_i\phi \stackrel{\text{def}}{\Longleftrightarrow} \phi \wedge \boldsymbol{B}_i\phi.$$

$\boldsymbol{K}_i\phi$ is read as " the agent i knows ϕ".

11.5.4. Proposition. (More properties of the knowledge operator) *The following axioms about the knowledge operator are sound for the class of ALX3 models.*

(KK) $\boldsymbol{K}_i\phi \wedge \boldsymbol{K}_i(\phi \rightarrow \psi) \rightarrow \boldsymbol{K}_i\psi$.
(TK) $\boldsymbol{K}_i\phi \rightarrow \phi$.

(K∧) $\boldsymbol{K}_i\phi \wedge \boldsymbol{K}_i\psi \leftrightarrow \boldsymbol{K}_i(\phi \wedge \psi)$.
(K∨) $\boldsymbol{K}_i\phi \vee \boldsymbol{K}_i\psi \rightarrow \boldsymbol{K}_i(\phi \vee \psi)$.
(4K) $\boldsymbol{K}_i\phi \rightarrow \boldsymbol{K}_i\boldsymbol{K}_i\phi$.

PROOF. Straightforward. □

11.6 Application of ALX3

In this section, we discuss the application of ALX. As argued in the introduction, we develop ALX as formal language for social science theories, especially for theories of organizations. H. A. Simon's bounded rationality is an important notion in theories of organizations. Goals which are defined in terms of preferences offer a natural and powerful tool in the formalization of theories concerning agents with bounded rationality. In [Masuch&Huang 1994], we use ALX to formalize J. D. Thompson's *Organization in Action*. In the application of ALX, we consider several problems which are discussed in the following.

11.6.1 Second Order Quantifier on Preference Formulas

In the applications of ALX, users may want to use some second order quantifier on preference formulas. The most-preferred operator is one of examples. One of the typical second order formulas is like:

$$\forall\psi(\psi\boldsymbol{P}_i\phi \rightarrow \Phi(\psi)).$$

where $\Phi(\psi)$ is a formula in which ψ appears. We call a second order formula of the above form a *second order preference formula*. In the following, we will focus on the problem how we can use second order preference formulas in the application.

For a formula set Γ, we use

$$\Gamma \vdash \forall\psi(\psi\boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$$

to denote "$\Gamma \vdash \psi\boldsymbol{P}_i\phi \rightarrow \Phi(\psi)$, for all formula ψ", which equally means that "$\Gamma \vdash \Phi(\psi)$, for all ψ such that $\Gamma \vdash \psi\boldsymbol{P}_i\phi$".

Furthermore, for formula sets Γ and Σ, we use the notation

$$\Gamma \vdash \forall^{\Sigma}\psi(\psi\boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$$

to denote "$\Gamma \vdash \psi\boldsymbol{P}_i\phi \rightarrow \Phi(\psi)$, for all formulas $\psi \in \Sigma$", which equally means that "$\Gamma \vdash \Phi(\psi)$, for all $\psi \in \Sigma$ such that $\Gamma \vdash \psi\boldsymbol{P}_i\phi$".

Therefore, we have:

$$\Gamma \vdash \forall^{FML}\psi(\psi\boldsymbol{P}_i\phi \rightarrow \Phi(\psi)) \iff \Gamma \vdash \forall\psi(\psi\boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$$

where FML is the set of all formulas.

Specially, if Σ is a finite set, say, $\Sigma = \{\psi_1, ..., \psi_n\}$, then,

$$\Gamma \vdash \forall^{\Sigma}\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi)) \iff \Gamma \vdash (\psi_1 \boldsymbol{P}_i\phi \rightarrow \Phi(\psi_1)) \wedge ... \wedge (\psi_n \boldsymbol{P}_i\phi \rightarrow \Phi(\psi_n)).$$

Note that the formula $(\psi_1 \boldsymbol{P}_i\phi \rightarrow \Phi(\psi_1)) \wedge ... \wedge (\psi_n \boldsymbol{P}_i\phi \rightarrow \Phi(\psi_n))$ is an ALX well-formed formula. Therefore, if we assume that whenever users say something in a second order preference formulas, users always mean the second order formulas with respect to a finite first order formulas set (we call this assertion the *finiteness assumption*.), then users would have no real problem to use the second order formula in the application of ALX, since users can always substitute the second order preference formula by a finite set of first order formulas.

Now, the problem becomes this: does the finiteness assumption make sense? In other words, whenever users use a second order preference formula to formalize a social theory, can users always find a finite first order set to substitute for the second order formula? In the following, we will examine the problem in details.

Given a social theory, assume that users formalize the theory by a finite formula set Γ. The set Γ is called a premise set. Sometimes, the user may think ALX is not powerful enough to formalize the theory, therefore, she may want to use some extra formula concerning second preference formulas to enlarge the formalization, called the new set Γ^+. Moreover, if user can find a finite formula set which is *reasonable in some sense*, she always can substitute the second order formuals by the finite set to make the new set Γ^+ ALX well-formed. Reasonableness here can mean anything, which only depends on users. However, in the following, we will suggest a definition of reasonablity

11.6.1. Definition. (Reasonable set) *Given a premise set Γ and a second order preference formula $\forall\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$, a formula set Σ is said to be a reasonable set with respect to Γ and the second order preference formula, if and only if it satisfies the following conditions:*
(i) Σ is finite, and
(ii) for any ψ such that $\Gamma \vdash \psi \boldsymbol{P}_i\phi$, there exists a formula $\psi' \in \Sigma$ such that $\vdash (\psi' \leftrightarrow \psi)$.

The proposed definition of reasonable sets above is intuitive, since the condition (ii) actually says that

$$\Gamma \vdash \forall^{\Sigma}\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi)) \iff \Gamma \vdash \forall\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi)).$$

Note that, since Σ is finite, $\forall^{\Sigma}\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$ is ALX well-formed. The above statement actually says that we can use the finite reasonable set Σ to substitute for the second order preference formula without loss of the original meaning of the second order formula.

11.6.1. Proposition. *For any premises set Γ and a second order preference formula $\forall\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$, if a set Σ is a reasonable set with respect to Γ and the second order preference formula, then*

$$\Gamma \vdash \forall^{\Sigma}\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi)) \iff \Gamma \vdash \forall\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi)).$$

PROOF.
($\Rightarrow$)
$\Gamma \vdash \forall^{\Sigma}\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$
$\Rightarrow \Gamma \vdash \Phi(\psi)$ for all $\psi \in \Sigma$ such that $\Gamma \vdash \psi \boldsymbol{P}_i\phi$ (Definition of $\forall^{\Sigma}\psi$)

For any formula ψ such that $\Gamma \vdash \psi \boldsymbol{P}_i\phi$.
$\Gamma \vdash \psi \boldsymbol{P}_i\phi$
$\Rightarrow \Gamma \vdash (\psi \boldsymbol{P}_i\phi)$ and there exists a $\psi' \in \Sigma$ such that $\vdash (\psi' \leftrightarrow \psi)$ (Condition (ii))
$\Rightarrow \Gamma \vdash (\psi' \boldsymbol{P}_i\phi)$ and $\psi' \in \Sigma$ and $\vdash \psi' \leftrightarrow \psi$ (Substitution rule)
$\Rightarrow \Gamma \vdash \Phi(\psi')$ and $\vdash \psi' \leftrightarrow \psi$ (Known fact)
$\Rightarrow \Gamma \vdash \Phi(\psi)$ (Substitution rule)
Therefore, we have:
$\Gamma \vdash \Phi(\psi)$ for all ψ such that $\Gamma \vdash \psi \boldsymbol{P}_i\phi$
$\Rightarrow \Gamma \vdash \forall\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$ (Definition of $\forall\psi$)

($\Leftarrow$)
$\Gamma \vdash \forall\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$
$\Rightarrow \Gamma \vdash \Phi(\psi)$ for all formula ψ such that $\Gamma \vdash \psi \boldsymbol{P}\phi$ (Definition of $\forall\psi$)
$\Rightarrow \Gamma \vdash \Phi(\psi)$ for all formula $\psi \in \Sigma$ such that $\Gamma \vdash \psi \boldsymbol{P}_i\phi$ ($\Sigma \subseteq FML$)
$\Rightarrow \Gamma \vdash \forall^{\Sigma}(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$ (Definition of $\forall^{\Sigma}\psi$) □

So our problem reduces to the following question: given a finite premise set Γ, how can we find a reasonable set Σ with respect to Γ? In the following, we offer several theorems about the reasonable sets. Moreover, we show that the reasonable sets always exist for propositional ALX logic. Furthermore, we show that in many applications the reasonable sets for first order ALX exist as well. First, we need the following definitions:

11.6.2. Definition. (Preference formula) *A preference formula is a formula with a form like $\phi \boldsymbol{P}_i\psi$. The set of preference formulas PF is defined by the following rule:*

$$\phi, \psi \in FML, i \in AGENT \Rightarrow \phi \boldsymbol{P}_i\psi \in PF.$$

11.6.3. Definition. (Logical closure) *For any formula set Γ, define the logical closure of Γ as follows:*

$$Cn(\Gamma) \stackrel{\text{def}}{\Longleftrightarrow} \{\phi \in FML : \Gamma \vdash \phi\}.$$

11.6.4. Definition. (Formula set under the logical equivalence) *First we define an equivalence relation $\approx$ on formulas as follows:*

$$\phi \approx \psi \stackrel{\text{def}}{\Longleftrightarrow} \vdash \phi \leftrightarrow \psi.$$

Define the equivalence class under the relation $\approx$ as usual, namely,

$$[\phi] = \{\psi \in FML : \phi \approx \psi\}.$$

For each equivalence class $[\phi]$*, we pick up a formula* ϕ *as the representative of the class. Therefore, for any formula set* Γ*, the set of* Γ *under the logical equivalence is defined as :*

$$\Gamma/_{\approx} \stackrel{\text{def}}{\Longleftrightarrow} \{\phi \in FML : \phi \text{ is the representative of the class } [\psi] \text{ for each } \psi \in \Gamma\}.$$

11.6.5. Definition. (Direct subformula of preference) *For a preference formula set* Σ*, we define the direct subformula set of* Σ *as:*

$$DSS(\Sigma) = \{\psi \in FML : \psi \boldsymbol{P}_i \phi \in \Sigma \text{ or } \phi \boldsymbol{P}_i \psi \in \Sigma\}.$$

In the following, we first consider the case of propositional ALX; then we consider whether or not the result can be generalized for the first order case. With regards to the propositional case, we have the following theorem.

11.6.1. Theorem. (Existence theorem of reasonable set for propositional ALX) *For any finite premise set* Γ *and any second order preference formula* $\forall\psi(\psi \boldsymbol{P}_i \phi \rightarrow \Phi(\psi))$*, the direct subformula set of preference formulas in the logical closure of* Γ *under the logical equivalence is a reasonable set;* $DSS(Cn(\Gamma) \cap PF)/_{\approx}$ *is a reasonable set with respect to* Γ *and the second order preference formula.*

PROOF. Let $\Sigma = DSS(Cn(\Gamma) \cap PF)/_{\approx}$.

i) We have to prove that Σ is finite.
Since the premise set Γ is finite, clearly the preference formulas in the subformula set of Γ is finite, namely, the set $Subformula(\Gamma) \cap PF$ is finite. Because we only consider preference formulas, only preference axioms (i.e., (CEP),(IRE), (N), and (TR)) and the inference rules (MP) and (SUBP) can yield new preference formulas in the derivation. Moreover, note that axioms (N) and (IRE) yield no new preference formulas. Furthermore, since we consider the set under the logical equivalence, the substitution rule (SUBP) yields no new preference formulas as well. Therefore, actually only axioms (CEP), (TR) under modes ponus (MP) contribute of new preference formulas. However, the transitivity axiom on finite preference formulas yields a finite set, and the conjunction expansion principle (CEP) yields only one more preference statement (Thanks to the fact $\Gamma \vdash \phi \boldsymbol{P}_i \psi \Rightarrow \Gamma \vdash (\phi \wedge \neg\psi) \boldsymbol{P}_i (\psi \wedge \neg\phi) \Rightarrow \Gamma \vdash ((\phi \wedge \neg\psi) \wedge \neg(\psi \wedge \neg\phi)) \boldsymbol{P}_i ((\psi \wedge \neg\phi) \wedge \neg(\phi \wedge \neg\psi))$ but $((\phi \wedge \neg\psi) \wedge \neg(\psi \wedge \neg\phi)) \approx (\phi \wedge \neg\psi)$ and $((\psi \wedge \neg\phi) \wedge \neg(\phi \wedge \neg\psi)) \approx (\psi \wedge \neg\phi)$, Therefore, the derivation stops at the second step.).

Therefore, we conclude Σ is finite.

ii) We have to prove that the condition (ii) of a reasonable set is satisfied.

For any formula ψ such that $\Gamma \vdash \psi \boldsymbol{P}_i \phi$,
$\Gamma \vdash \psi \boldsymbol{P}_i \phi$
$\Rightarrow \psi \boldsymbol{P}_i \phi \in Cn(\Gamma) \cap PF$ (Definitions of Cn and PF)
$\Rightarrow \psi \in DSS(Cn(\Gamma) \cap PF)$ (Definition of DSS)

$\Rightarrow$ there exists a $\psi' \in DSS(Cn(\Gamma) \cap PF)/_{\approx}$ such that $\vdash \psi' \leftrightarrow \psi$ (Definition of $\approx$)
$\Rightarrow$ there exists a $\psi' \in \Sigma$ such that $\vdash \psi' \leftrightarrow \psi$ (Definition of Σ)

Therefore, Σ is a reasonable set. □

Therefore, for any finite premise set Γ (in propositional ALX, i.e., ALX1), the reasonable set always exists. Furthermore, since the reasonable sets are finite, we can have an algorithm to create the reasonable set automatically. Whenever users have a premise set Γ^+ in which there exists some second order preference formulas, users can use the algorithm to translate Γ^+ into a set of ALX well-found formulas.

Now, we consider the case of first order ALX, i.e., ALX3. The above theorem cannot be simply generalized for first order case, since in the first order ALX, we allow to use countable constants. There is a danger that the instance set of the direct subformula set of the preference formula under the logical equivalence still is infinite.

In order to capture reasonable sets for the first order ALX, one of the natural considerations is to use the closed domain assumption (i.e., if an element is not mentioned in the premise set, then the element does not exist); this assumption is used frequently in artificial intelligence and computer science.

In the following, we consider another alternative, which also is natural in many applications.

11.6.6. Definition. (Finite branching property) *For a formula set Γ and a second order preference formula $\forall\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$, Γ is said to have the finite branching property with respect to the second order preference formula iff the set $\{\psi \in FML : \Gamma \vdash \psi \boldsymbol{P}_i\phi\}/_{\approx}$ is finite, namely, there exist only finitely many formulas, say ψ, under the logical equivalence such that $\Gamma \vdash \psi \boldsymbol{P}_i\phi$.*

11.6.2. Theorem. (Existence theorem of reasonable sets for the first order ALX) *For a formula set Γ and a second order preference formula $\forall\psi(\psi \boldsymbol{P}_i\phi \rightarrow \Phi(\psi))$, if Γ has the finite branching property with respect to the second order preference formula, then the reasonable set exists.*

PROOF. Let $\Sigma = \{\psi \in FML : \Gamma \vdash \psi \boldsymbol{P}_i\phi\}/_{\approx}$. It is straightforward to see that Σ is the reasonable set we need, by the definitions. First, from the definition of the finite branching property, we have that Σ is finite. Moreover, from the definition of $\approx$, we know that for any ψ such that $\Gamma \vdash \psi \boldsymbol{P}_i\phi$, there exists a $\psi' \in \Sigma$ such that $\vdash \psi' \leftrightarrow \psi$. So the condition (ii) of the reasonable set is satisfied. Therefore, Σ is a reasonable set with respect to Γ and the second order preference formula. □

We summarize the discussion in the following:
1) Users would have no real problems when they use second order preference formulas if they can find a finite first order formula set to substitute the second order formulas without the loss of the original meaning of the second order formula.
2) Reasonable sets are the finite formula sets which users can substitute in the second order preference formula without the loss of the original meaning of the second order

formula.
3) For any finite premise set in propositional ALX, the reasonable set always exists.
4) For any premise set in first order ALX, if the premise set has the finite branching property, then the reasonable set exists as well.

11.6.2 Agents and Accessible States

In the applications, users may want to express that some agent is one of actors for an action. Users can use a predicate $AGT(i, a)$ to denote that "agent i is one of the actors of action a". The agent theory can be expressed as follows:

(AG1) $AGT(i, a_i)$.
(AG2) $AGT(i, (a; b)) \leftrightarrow AGT(i, a) \vee AGT(i, b)$.
(AG3) $AGT(i, (a \cup b)) \leftrightarrow AGT(i, a) \vee AGT(i, b)$.

Here we define that agent i is one of agents of action $a \cup b$ if and only if the agent i is one of agents of either a or b, since we need the property which says that there exists at least one agent for any action. Otherwise we may introduce a stronger definition for that case as follows:

(AG3*) $AGT(i, a \cup b) \leftrightarrow AGT(i, a) \wedge AGT(i, b)$.

Sometimes users may want to express that some agent is the only agent of the action a. Using *Only-agent*(i, a) to denote that "agent i is the only agent of the action a", which can be defined as:

$$\textit{Only-agent}(i, a) \stackrel{\text{def}}{\Longleftrightarrow} AGT(i, a) \wedge \forall j(AGT(j, a) \rightarrow eq(i, j)).$$

Where the predicate $eq(i, j)$ means that the agent i is equal to the agent j. Note that we have not yet introduced the equality in the object language of ALX3. However, in the applications, we may add the following equality theory to capture partial meaning of the equality:

(Reflexivity) $\forall x\ eq(x, x)$.
(Symmetry) $\forall xy(eq(x, y) \rightarrow eq(y, x))$.
(Transitivity) $\forall xyz(eq(x, y) \wedge eq(y, z) \rightarrow eq(x, z))$.

We use $\boldsymbol{A}_i\phi$ to denote that the state ϕ is accessible for agent i via some action a. However, there are different readings regarding accessibility. We have the different alternatives:

(1) $\boldsymbol{A}^1_i\phi \stackrel{\text{def}}{\Longleftrightarrow} \langle a_i\rangle\phi$.
(2) $\boldsymbol{A}^2_i\phi \stackrel{\text{def}}{\Longleftrightarrow} [a_i]\phi$.
(3) $\boldsymbol{A}^3_i\phi \stackrel{\text{def}}{\Longleftrightarrow} (\langle a\rangle\phi \wedge AGT(i, a))$.
(4) $\boldsymbol{A}^4_i\phi \stackrel{\text{def}}{\Longleftrightarrow} ([a]\phi \wedge AGT(i, a))$.

(5) $\boldsymbol{A}_i^5\phi \overset{\text{def}}{\Longleftrightarrow} (\langle a\rangle\phi \wedge \mathit{Only\text{-}agent}(i,a))$.
(6) $\boldsymbol{A}_i^6\phi \overset{\text{def}}{\Longleftrightarrow} ([a]\phi \wedge \mathit{Only\text{-}agent}(i,a))$.

The relation between those different readings of accessibility is:

$$\Sigma \vdash \boldsymbol{A}_i^6\phi \Rightarrow \Sigma \vdash \boldsymbol{A}_i^4\phi \Rightarrow \Sigma \vdash \boldsymbol{A}_i^2\phi.$$

$$\Sigma \vdash \boldsymbol{A}_i^5\phi \Rightarrow \Sigma \vdash \boldsymbol{A}_i^3\phi \Rightarrow \Sigma \vdash \boldsymbol{A}_i^1\phi.$$

Where Σ is a formula set.

11.6.3 Goals

Axioms about Goals

In general, a state ϕ is a goal for agent i if ϕ is preferred (in some sense) to other states and accessible (or at least not believed to be inaccessible) for agent i. At this point, we do not want to give a unique definition for goals, since the notion is intrinsically ambivalent. In the following, we consider some plausible axioms for goals, consider some alternative definitions, and check finally which axioms are satisfied by which definitions.

In the following, we use $\boldsymbol{G}_i\phi$ to denote "the state ϕ is a goal for agent i".

Axioms for goals
(G1) $\neg\boldsymbol{G}_i\top$.
(A tautology is never a goal.)

(G2) $\boldsymbol{G}_i\phi \rightarrow \neg\boldsymbol{G}_i\neg\phi$.
(Goals are not contradictory.)

Axioms for goals and preferences
(GP1) $\boldsymbol{G}_i\phi \rightarrow \neg(\neg\phi\boldsymbol{P}_i\phi)$.
(Goals are not "bad".)

(GP1*) $\boldsymbol{G}_i\phi \rightarrow \phi\boldsymbol{P}_i\neg\phi$.
(Goals are "good". It is easy to see that (GP1*) implies (GP1)).

(GP2) $\phi\boldsymbol{P}_i\psi \wedge \boldsymbol{G}_i\phi \rightarrow \neg\boldsymbol{G}_i\psi$.
(If agent prefers ϕ to ψ, and ϕ is a goal for agent i, then ψ is not a goal for agent i.)

(GP3) $\boldsymbol{G}_i\phi \wedge \boldsymbol{G}_i\psi \wedge (\phi\wedge\psi)\boldsymbol{P}_i\neg(\phi\wedge\psi) \rightarrow \boldsymbol{G}_i(\phi\wedge\psi)$.
(If both ϕ and ψ are goals for agent i, and agent i thinks the conjunction is good, then the conjunction is a goal as well.)

Axioms for goals and beliefs

(GB1) $\boldsymbol{G}_i\phi \rightarrow \boldsymbol{B}_i\boldsymbol{G}_i\phi$.
(Agents know their goals)

(GB2) $\boldsymbol{G}_i\phi \rightarrow \neg\boldsymbol{B}_i\boldsymbol{G}_i\neg\phi$.
(If agent i has goal ϕ, then she does not believe that $\neg\phi$ is one of her goals.)

(GB3) $\boldsymbol{B}_i\boldsymbol{G}_i\phi \rightarrow \boldsymbol{G}_i\phi$.
(If agent i believes that ϕ is her goal, then ϕ is her goal.

Axioms about goal and accessibility
(GA1) $\boldsymbol{G}_i\phi \rightarrow \boldsymbol{A}_i\phi$.
(Goal must be accessible)

Axioms about goal, accessibility and preference
(GAP1) $\boldsymbol{G}_i\phi \wedge \psi\boldsymbol{P}_i\phi \rightarrow \neg\boldsymbol{A}_i\psi$.
(If agent i has a goal ϕ, and agent i prefers ψ to ϕ, then ψ must not be accessible for agent i.)

Definition of Goals

In the following, we offers different definitions for goals. In general, we consider the following four dimensions: (i) may-conflict vs conflict-free; (ii) preferred vs. most-preferred; (iii) accessibility vs accessibility-free; (iv) believed vs. believed-free.

We define:

(1) $\boldsymbol{G}_i^p\phi \stackrel{\text{def}}{\Longleftrightarrow} \phi\boldsymbol{P}_i\rho$.
(A goal is something preferred by the agent.)

(2) $\boldsymbol{G}_i^g\phi \stackrel{\text{def}}{\Longleftrightarrow} \phi\boldsymbol{P}_i\neg\phi$.
(A goal is something good for the agent.)

(3) $\boldsymbol{G}_i^a\phi \stackrel{\text{def}}{\Longleftrightarrow} \phi\boldsymbol{P}_i\rho \wedge \boldsymbol{A}_i\phi$.
(A goal is both preferred and accessible for the agent.)

(4) $\boldsymbol{G}_i^{ga}\phi \stackrel{\text{def}}{\Longleftrightarrow} \phi\boldsymbol{P}_i\neg\phi \wedge \boldsymbol{A}_i\phi$.
(a goal is something good and accessible for the agent.)

(5) $\boldsymbol{G}_i^b\phi \stackrel{\text{def}}{\Longleftrightarrow} \phi\boldsymbol{P}_i\neg\phi \wedge \forall\psi(\psi\boldsymbol{P}_i\phi \rightarrow \psi \wedge \neg\psi)$.

Here we define goals in terms of the most-preferred states, using a second order preference formula. Note that we have discussed the problem how to incorporate the second order preference formula in ALX before. This notion is expressed more

convenient as:

Most-preferred$_i\phi \stackrel{\text{def}}{\Longleftrightarrow} \boldsymbol{G}_i^b\phi$.

(6) $\boldsymbol{G}_i^{ba}\phi \stackrel{\text{def}}{\Longleftrightarrow}$ Most-preferred$_i\phi \wedge \boldsymbol{A}_i\phi$.
(A goal is a state both most-preferred and accessible for the agent.)
This definition sometimes is too strong, since the intersection of the most-preferred states and the accessible states may be empty.

(7) $\boldsymbol{G}_i^{bc}\phi \stackrel{\text{def}}{\Longleftrightarrow} \phi\boldsymbol{P}_i\neg\phi \wedge \forall\psi(\psi\boldsymbol{P}_i\phi \rightarrow \neg\boldsymbol{A}_i\psi)$.
(A goal is a state which is most-preferred with respect to all accessible states for the agent.)

(8) $\boldsymbol{G}_i^{bcB}\phi \stackrel{\text{def}}{\Longleftrightarrow} \phi\boldsymbol{P}_i\neg\phi \wedge \forall\psi(\psi\boldsymbol{P}_i\phi \rightarrow \neg\boldsymbol{B}_i\boldsymbol{A}_i\psi)$.
(A goal is a state which is most-preferred with respect to all believed accessible states for the agent.)

Considering the dimension "believed vs. believed-free", we have the following definitions:

(a) $\boldsymbol{G}_i^{B}\phi \stackrel{\text{def}}{\Longleftrightarrow} \boldsymbol{B}_i(\phi\boldsymbol{P}_i\rho)$.

(b) $\boldsymbol{G}_i^{Bg}\phi \stackrel{\text{def}}{\Longleftrightarrow} \boldsymbol{B}_i(\phi\boldsymbol{P}_i\neg\phi)$.

(c) $\boldsymbol{G}_i^{Ba}\phi \stackrel{\text{def}}{\Longleftrightarrow} \boldsymbol{B}_i(\phi\boldsymbol{P}_i\rho \wedge \boldsymbol{A}_i\phi)$.

(d) $\boldsymbol{G}_i^{Bga}\phi \stackrel{\text{def}}{\Longleftrightarrow} \boldsymbol{B}_i(\phi\boldsymbol{P}_i\neg\phi \wedge \boldsymbol{A}_i\phi)$.

(e) $\boldsymbol{G}_i^{Bb}\phi \stackrel{\text{def}}{\Longleftrightarrow} \boldsymbol{B}_i$Most-preferred$_i\phi$.

(f) $\boldsymbol{G}_i^{Bba}\phi \stackrel{\text{def}}{\Longleftrightarrow} \boldsymbol{B}_i($Most-preferred$_i\phi \wedge \boldsymbol{A}_i\phi)$.

(g) $\boldsymbol{G}_i^{Bbc}\phi \stackrel{\text{def}}{\Longleftrightarrow} \boldsymbol{B}_i(\phi\boldsymbol{P}_i\neg\phi) \wedge \forall\psi(\psi\boldsymbol{P}_i\phi \rightarrow \neg\boldsymbol{A}_i\psi)$.

(h) $\boldsymbol{G}_i^{BbcB}\phi \stackrel{\text{def}}{\Longleftrightarrow} \boldsymbol{B}_i(\phi\boldsymbol{P}_i\neg\phi) \wedge \forall\psi(\psi\boldsymbol{P}_i\phi \rightarrow \neg\boldsymbol{B}_i\boldsymbol{A}_i\psi)$.

Goal Analysis

We define goals in terms of preference, in order to avoid the counter-intuitive properties of goals. In other action logics [Rao&Georgeff 1991] goal operators are introduced as syntactic primitives acting like universal modalities. As a consequence, these logics have the necessitation rule for goals (if α is a theorem, then α is a goal), and the closure of goals under logical implication (if α is a goal, and $\alpha \rightarrow \beta$ is a theorem, then β must be a goal). The necessitation rule and the deductive closure of goals have fairly severe counterintuitive implications. For example, if tooth-ache

is always a consequence of having one's teeth restored, then tooth-ache appears as a goal itself. Also, it does not make sense to treat tautologies as goals, as the necessitation rule would require. Much recent work in action logic has gone into systems that are trying to avoid these consequences by introducing an array of goal-related notions [Cohen&Levesque 1987, Cohen&Levesque 1990, Rao&Georgeff 1991]. Unfortunately, these complications bring in other, or additional, counterintuitive effects of goals. For example, in Cohen and Levesque's logic [Cohen&Levesque 1987, Cohen&Levesque 1990], it is a theorem that if an agent believes that a fact holds, then the fact becomes a goal for this agent. Rao and Georgeff's recent paper [Rao&Georgeff 1991] avoids both necessitation and logical closure for certain epistemically qualified goals (agents need not adopt as goals what they *believe* to be *inevitably always* true, and they need not to adopt ψ as a goal if they *believe* $\phi \rightarrow \psi$ to be *inevitably always* true and if they have ϕ as a goal). But in order to obtain these results, Rao and Georgeff have to make another counterintuitive assumption. For example, they must assume that any believe-accessible world contains a goal. ALX can avoid both the necessitation rule and the deductive closure of goals by much simpler means, thanks to the fact we need not require monotonicity for the preference operator.

We want to know which axioms are satisfied by the various goal definitions:

Goals	**Axioms**
$\boldsymbol{G}_i^{p}$	(G1)
$\boldsymbol{G}_i^{g}$	(G1), (G2), (GP1)
$\boldsymbol{G}_i^{a}$	(G1), (GA1)
$\boldsymbol{G}_i^{ga}$	(G1), (G2), (GA1), (GP1)
$\boldsymbol{G}_i^{b}$	(G1),(G2), (GP1),(GP2)
$\boldsymbol{G}_i^{ba}$	(G1),(G2),(GP1), (GP2), (GA1), (GAP1)
$\boldsymbol{G}_i^{bc}$	(G1),(G2),(GP1), (GP2), (GA1), (GAP1)
$\boldsymbol{G}_i^{bcB}$	(G1),(G2),(GP1), (GP2), (GA1), (GAP1)
$\boldsymbol{G}_i^{B}$	(G1), (GB1)
$\boldsymbol{G}_i^{Bg}$	(G1), (G2), (GP1), (GB1), (GB2)
$\boldsymbol{G}_i^{Ba}$	(G1), (GA1)
$\boldsymbol{G}_i^{Bga}$	(G1), (G2), (GA1), (GP1), (GB1), (GB2)
$\boldsymbol{G}_i^{Bb}$	(G1),(G2), (GP1),(GP2)
$\boldsymbol{G}_i^{Bba}$	(G1),(G2),(GP1), (GP2), (GA1), (GAP1), (GB1), (GB2)
$\boldsymbol{G}_i^{Bbc}$	(G1),(G2),(GP1), (GP2), (GA1), (GAP1), (GB1), (GB2)
$\boldsymbol{G}_i^{BbcB}$	(G1),(G2),(GP1), (GP2), (GA1), (GAP1), (GB1), (GB2)

The relations between the goals are shown in the figure 11.1. An arrow denotes logical implication. For instance, $\boldsymbol{G}_i^{g} \rightarrow \boldsymbol{G}_i$ means that $\Sigma \vdash \boldsymbol{G}_i^{g}\phi \Rightarrow \Sigma \vdash \boldsymbol{G}_i\phi$ where Σ is a formula set.

The relations for believed goals have a similar structure.

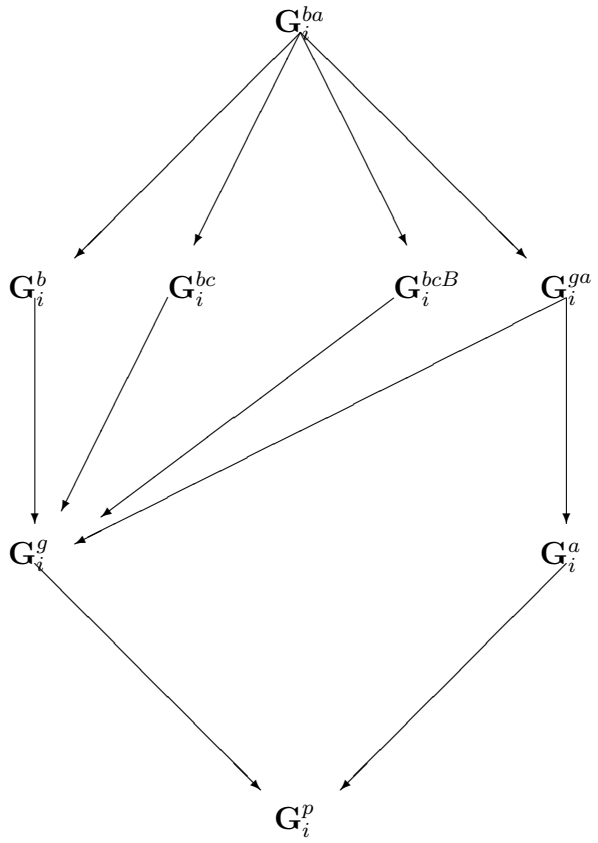


Figure 11.1: Relations between goals

11.6.4 Power

The notions of control, influence, dependence, and power play important parts in reasoning about organizations. We define all of those notions in terms of a fundamental modal operation "bring-about". In the following, we use $\textit{Can-bring-about}_i\phi$ to denote that the agent i can bring about a state ϕ.

$$\textit{Can-bring-about}_i\phi \stackrel{\text{def}}{\Longleftrightarrow} \exists a([a_i]\phi \wedge \neg\Box\phi).$$

(The agent i can bring about ϕ iff there exists a primitive action a such that the agent i is the agent of the action a and after the action a, ϕ will hold, and ϕ does not hold necessarily.)

Similar to the action accessibility operator, there exist several different readings for "can bring about". Therefore, we have:
$\textit{Can-bring-about}_i^o\phi \stackrel{\text{def}}{\Longleftrightarrow} (\textit{Only-agent}(i,a) \wedge [a]\phi \wedge \neg\Box\phi)$.

(The agent i can bring about ϕ iff there exists an action a such that the agent i is the only agent of the action a and after the action a, ϕ will hold, and ϕ does not hold necessarily.)

Sometimes we may need a weaker notion of "can-bring-about", in which the agent i is one of the agents for an action.

$$\textit{Can-bring-about}_i^w\phi \stackrel{\text{def}}{\Longleftrightarrow} (AGT(i,a) \wedge ([a]\phi \wedge \neg\Box\phi)).$$

(The agent i can bring about ϕ iff there exists an action a such that the agent i is one of the agents of the action a and after the action a, ϕ will hold, and ϕ does not hold necessarily.)

It is easy to see that the following properties hold.

(CB1) $\vdash \neg\textit{Can-bring-about}_i\top$
(CB2) $\vdash \textit{Can-bring-about}_i^o\phi \Rightarrow\vdash \textit{Can-bring-about}_i^w\phi \Rightarrow\vdash \textit{Can-bring-about}\phi$.

In [Kuhn 1974], Kuhn distinguishes intellectual influence and moral influence. Intellectual influence is defined as the ability, through communication, to alter the detector of others so that certain things are no longer conceived or perceived as before, whereas control is defined as the power, i.e., the ability to bring about desired external states. From Kuhn's viewpoint, the main distinction between control and influence is that control focuses on the the change of external states, whereas influence focuses on the change of internal, or perceived, states. Therefore, according to Kuhn's approach, we define:

$$\textit{Can-control}_{i,j}\phi \stackrel{\text{def}}{\Longleftrightarrow} \textit{Can-bring-about}_i\,\textit{Can-bring-about}_j\phi.$$

(The agent i can control the agent j with respect to ϕ iff the agent i can bring about that the agent j can bring about ϕ.)

$\mathit{Can\text{-}influence}_{i,j}\phi \overset{\text{def}}{\Longleftrightarrow} \mathit{Can\text{-}bring\text{-}about}_i \boldsymbol{B}_j\phi.$

(The agent i can influence the agent j with respect to ϕ iff the agent i can bring about that the agent j believes ϕ.)

Moreover, we can formalize the notion of dependence as follows:

$\mathit{Dependent\text{-}on}_{i,j}\phi \overset{\text{def}}{\Longleftrightarrow} \boldsymbol{G}_i\phi \wedge \mathit{Can\text{-}control}_{j,i}\phi.$

(The agent i depends on the agent j with respect to ϕ iff the agent i has a goal ϕ but the agent j can control the agent i with respect to ϕ.)

$\mathit{Has\text{-}power}_{i,j}\phi \overset{\text{def}}{\Longleftrightarrow} \mathit{Dependent\text{-}on}_{j,i}\phi.$

(The agent i has power, relative to the agent j with respect to ϕ iff the agent j depends on the agent i with respect to ϕ.)

11.6.5 Cooperation and Coordination

Obviously the notions of cooperation and coordination play an important role in the formalization of social agents. in [Werner 1990], Eric Werner distinguishes two types of cooperation: negative cooperation and positive cooperation. *Negative cooperation* refers to the cooperation in which one agent does not take some action to achieve some shared goal, whereas *positive coordination* refers to the one in which one agent does take some action to achieve the shared goal. In the following, we define the positive cooperation.

First, we define the notion of shared goal as follows:

$\mathit{Shared\text{-}goal}_{i,j}\phi \overset{\text{def}}{\Longleftrightarrow} \boldsymbol{G}_i\phi \wedge \boldsymbol{G}_j\phi.$

The cooperation between two agents by action a with respect to ϕ is just the act of working together for a shared goal. So, we have:

$\mathit{Cooperation}^a_{i,j}\phi \overset{\text{def}}{\Longleftrightarrow} ([a]\phi \wedge AGT(i,a) \wedge AGT(j,a) \wedge \mathit{Shared\text{-}goal}_{i,j}\phi).$

We view the coordination as a special kind of cooperations where a single agent cannot achieve the goal.

$Cannot\text{-}achieve_i\phi \stackrel{\text{def}}{\Longleftrightarrow} \neg Can\text{-}bring\text{-}about_i\phi.$

$Coordination^a_{i,j}\phi \stackrel{\text{def}}{\Longleftrightarrow} Cooperation^a_{i,j}\phi \wedge Cannot\text{-}achieve_i\phi \wedge Cannot\text{-}achieve_j\phi.$

Following the general ideas we have developed above, we can capture more formal properties about the notions which are defined in this chapter and examine the theoretical implications in social science. However, the work is somehow a little bit out of this thesis's focus, since we want this part of this thesis more logically oriented than application oriented. It is no doubt that there is a lot of interesting, valueful and promising further work left behind the present work. Interesting further work is to study how to formally define more notions from social agent theory, using ALX3. Another interesting work is to use ALX3 logical machinery to capture more logical derivations in the formal theories we offer above, and think about more alternatives for these formal definitions, which would definitely offer us more insights about those crucial notions, both from a logical and a sociological perspectives.

11.7 Comparing ALX with Other Action Logics

There have been several proposed systems which are quite close to ALX logics. They are: Moore's dynamic epistemic logic, Cohen and Levesque's multi-agent action logic, Pörn's multi-agent logic, and Rao and Georgeff's action logic. In the following, first we briefly introduce those logics, then compare ALX with those logics with respect to the following two criteria:

(i) Expressibility: whether or not the logic has the abilities to define several modal operators such as belief, action, goal, etc.
(ii) Intuitivity: whether or not the logic can avoid several counterintuitive properties.

11.7.1 Other Action Logics

(1) **Moore's Dynamic Epistemic Logic**
In [Moore 1985], Moore proposes a formal theory of knowledge and action, which is actually a logic combining first order logic, an epistemic logic, and an action logic. The logic is called dynamic epistemic logic (J-J. Ch. Meyer's term [Meyer 1989]). Moreover, Moore defines the capability operator CAN in terms of the knowledge operator and the action operator. $CAN(i, act, \phi)$ means that "the agent i can achieve ϕ by performing action act. According to Moore, an agent i can achieve ϕ by performing action act if she knows what action act is, and she knows that ϕ would be true as a result of her performing action act. Moore expresses the fact by:

$$\forall i(\exists x K_i((x = act) \wedge [act_i]\phi) \rightarrow CAN(i, act, \phi)).$$

(2) **Cohen and Levesque's Multi-agent Action Logic**
Similar to Moore's work, in [Cohen&Levesque 1987], Cohen and Levesque propose a first order multi-agent logic, which combines first order logic, a belief logic, an action logic, and a temporal logic. One of the problems of Cohen and Levesque's multi-agent logic is that the goals in the logic suffer several counterintuitive properties since the goal accessibility relations are directly introduced in the semantic models. One of the features of the logic is that they define the intentions in terms of goals, and time and action operators so that the logic has the capability to express agents' action commitment. Cohen and Levesque argue that agents can distinguish between achievement goals and maintenance goals. Achievement goals are the goals the agent presently believes to be false; maintenance goals are the goals the agent already believes to be true. Achievement goals can be defined in the logic, since the logic has the time operators. Moreover, the logic can express that agents may eventually give up several achievement goals.

(3) **Pörn Multi-agent Logic**
In [Pörn 1989], Pörn proposes a multi-agent logic, called $DD'GOO'$, for formalizing theories of social orders. Actually Pörn's logic is one which combines a propositional logic, a possibility logic (i.e., ordinary modal logic for possibility and necessity operators), an action logic, and a deontic logic. Although in the logic there are no action names available, the operator $D'_i\phi$ with the meaning of that "for agent i's action it would be the case that ϕ" actually plays the same role in ALX logic like "$\exists a[a_i]\phi$". Moreover, although there is no directly goal operators in the logic, however, the operator $D_i\phi$ with the meaning of that "It is necessary for something which agent i does that ϕ" actually can be understood as an operator in ALX like "$\Box \boldsymbol{G}_i\phi$" or simply "$\boldsymbol{G}_i\phi$". One of the features of the logic is the use of deontic operators.

(4) **Rao and Georgeff's Action Logic**
In [Rao&Georgeff 1991], Rao and Georgeff propose a first order action logic which combines first order logic, a temporal logic, and a belief logic. Similar to Cohen and Levesque's approach, Rao and Georgeff introduce the goal accessibility relation directly in the semantics model. Therefore, goals in this logic suffer several counterintuitive properties. One of the features of this logic is that the intention accessibility relations are introduced directly in the semantics instead of defined in terms of other primitive operators like in Cohen and Levesque's logic.

11.7.2 Expressibility in Other Action Logics

We summerize the expressibilities among those logics in the following table:
In the table, p means the operator is primitive in the logic; d means that the operator is definable from the primitives operators.

	Moore	**Cohen and Levesque**	**Pörn**	**Rao and Georgeff**	**ALX3**
Multi-agents	yes	yes	yes		yes
First order logic	yes	yes		yes	yes
Belief operator		yes(p)		yes(p)	yes(p)
Knowledge operator	yes(p)	yes(d)			yes(d)
Action operator	yes(p)	yes(p)	yes(p)	yes(p)	yes(p)
Possibility operator		yes(p)	yes(p)	yes(p)	yes(d)
Preference operator					yes(p)
Conditional					yes(p)
Goal operator		yes(p)	yes(p)	yes(p)	yes(d)
Time operator		yes(p)		yes(p)	
Intention operator		yes(d)		yes(p)	
Deontic operator			yes(p)		

11.7.3 Avoidance of Counterintuitive Properties

We consider the following counterintuitive properties:

(A) Necessitation for goals
$\models \phi \Rightarrow \models \boldsymbol{G}_i\phi$.
(Something is tautological, then it is a goal.)

(B) Closure for goals
$\models \phi \rightarrow \psi \Rightarrow \models \boldsymbol{G}_i\phi \rightarrow \boldsymbol{G}_i\psi$.
(If $\phi \rightarrow \psi$ is a tautology, then having a goal ϕ implies having a goal ψ.)

(C) Closure under logical implication
$\models \boldsymbol{G}_i\phi \wedge \boldsymbol{G}_i(\phi \rightarrow \psi) \rightarrow \boldsymbol{G}_i\psi$.

(D) Beliefs imply goals
$\models \boldsymbol{B}_i\phi \rightarrow \boldsymbol{G}_i\phi$.

(E) Closure for expected consequences
$\models \boldsymbol{G}_i\phi \wedge \boldsymbol{B}_i(\phi \rightarrow \psi) \rightarrow \boldsymbol{G}_i\psi$.
(Actually, $(C) + (D) \Rightarrow (E)$.)

(F) Closure for necessary consequences
$\models \boldsymbol{G}_i\phi \wedge \boldsymbol{B}_i\Box(\phi \rightarrow \psi) \rightarrow \boldsymbol{G}_i\psi$.
(where $\Box$ is the necessity operator.)

We summerize the properties in the table below:
In the table "undefined" means that the belief operator $\boldsymbol{B}_i$ is undefined in the logic. Whence makes no sense to consider the properties (D), (E), and (F).

	Cohen and Levesque	Pörn	Rao and Georgeff	ALX3
(A)	yes	yes	yes	no
(B)	yes	yes	yes	no
(C)	yes	yes	yes	no
(D)	yes	undefined	no	no
(E)	yes	undefined	no	no
(F)	yes	undefined	no	no

11.7.4 Comparison by Examples

Example 1: Pörn's Social Order Theory
In[Pörn 1989], Pörn focus on the following operators to formalize the social order theory:

$E_i\phi$: the agent i brings about that ϕ.
$F_i\phi$: the agent i lets it be the case that ϕ.

Pörn defines those operators as:
$E_i\phi \overset{\text{def}}{\Longleftrightarrow} D_i\phi \wedge \neg D'_i\phi$.
$F_i\phi \overset{\text{def}}{\Longleftrightarrow} \neg D_i\neg\phi \wedge \neg D'_i\phi$.

Following the same idea, we can use ALX to formalize those operators as follows:
$E_i\phi \overset{\text{def}}{\Longleftrightarrow} \Box \boldsymbol{G}_i\phi \wedge \langle a_i\rangle\neg\phi$.
$F_i\phi \overset{\text{def}}{\Longleftrightarrow} \neg\Box \boldsymbol{G}_i\neg\phi \wedge \langle a_i\rangle\neg\phi$.

Moreover, Pörn introduces an operator $E_i(\phi,\psi)$ to denote "By bringing it about that ϕ, the agent i brings about that ψ". Pörn defines the operator as :

$$E_i(\phi,\psi) \overset{\text{def}}{\Longleftrightarrow} E_i\phi \wedge (\phi \Rightarrow \psi).$$

Where $\phi \Rightarrow \psi$ means that "ϕ leads to ψ". However, Pörn does not mention how he can define the "leads to" operator in his logic. Actually, ALX does better at this point, since ALX has the conditional operator, which can play a role as the "leads to" operator. So, in ALX, we can define:

$$E_i(\phi,\psi) \overset{\text{def}}{\Longleftrightarrow} E_i\phi \wedge (\phi \rightsquigarrow \psi).$$

As a result, we have:

$$E_i(\phi,\psi) \iff \Box \boldsymbol{G}_i\phi \wedge \langle a_i\rangle\neg\phi \wedge (\phi \rightsquigarrow \psi).$$

Example 2: The Egocentric Agent

[van der Hoek et al., 1993] offers the following egocentirc agent example: "If agent i knows that he will feel better after helping his neighbour and he knows that he is able to help him, he will do so, otherwise he will do nothing." The statement can be formalized by ALX as follows:

$$\boldsymbol{K}_i[help - neighbour_i]better(i) \wedge \boldsymbol{K}_i\boldsymbol{A}_i helped(neighbour) \leftrightarrow \boldsymbol{G}_i helped(neighbour).$$

Without using ALX, van der Hoek et al. formalize the statement by their logic of capabilities as follows:

$$w \rightarrow [do_i(if\, \boldsymbol{K}_i[do_i(h)]b \wedge \boldsymbol{K}_i\boldsymbol{A}_i h \text{ then } h \text{ else skip } fi)](w \vee b).$$

11.8 Final Remarks

11.8.1 Conclusions

In general, ALX logics differ from the existing multi-agent logics in the following respects:

1) ALX has a preference logic as its subsystem, which makes it more convenient for users to formalize social theories. Moreover, goals can be defined in terms of preferences, which can avoid several counter intuitive properties of goals.
2) ALX introduces a dynamic logic and a conditional logic as its subsystems, which offers a powerful tool to formalize actions, minimal change actions and their changes.
3) ALX logics enjoy several nice logical properties. ALX logics are complete with respect to their semantic models. In particular, ALX1 logic is decidable (i.e., the problem of the satisfiability of ALX1 formula is decidable), whereas ALX3 allows quantifiers over actions and agents, which makes the language more expressive and more flexible for many applications.

Doubtless there is a lot of further work to be done for ALX. Some of them are:
1) Introducing time operators, since many notions in social theories have close relation with the notion of time.
2) Consider the relation with deontic operator, since ALX has preference operator, by which we can define the goodness and badness operators. Moreover, we believe that there is a possibility to define several deontic operators.
3) ALX logics as decision logics. As a matter of fact, our definition for goals in terms of preference and action accessibility is a simplified case of a decision problem (i.e., a decision problem concerning single agent under certainty). There is indeed a possibility to develop ALX logics towards decision logics.
4) Develop a formal theory of power in social theories.

11.8.2 Bounded Rationality

We have tried to incorporate important elements of bounded rationality into ALX. It has not been very difficult to transpose Simon's original conceptualization of bounded

rationality into an action logic. However, notions of bounded rationality have proliferated since, and we cannot claim to cover all relevant aspect of bounded rationality as it might be understood now.

The basic message of the bounded rationality is quite simple: don't forget the limits of human information-processing capacity. Yet it is one thing to recognize the abstract existence of these limits, and another to find out where these limits are drawn. In the first case, one has to make sure that omniscience claims or omnipotence claims are avoided. In the second case, one has to identify which information is processed, and how. No logic would be able to fully answer the second question, since it is to a large extent an empirical one, but it does have some general aspects, that we did not address in this thesis. Furthermore, ALX has no way to model the consequences (and constraints) of *search*, i.e., the fact that knowledge might not be available all at once but must be actively acquired. In particular, the notion of *satisfying* (a given aspiration level) is not incorporated in ALX, since this notion involves search. Satisfying is theoretically justified by the fact that time constraints do not allow the agent to evaluate all conceivable alternatives at once, so that the agent may stop *searching* once an alternative meets prior aspirations. Incorporating search explicitly in the logic seems to require introducing "information" as a distinct object to the logic. This, in turn, seems to require partial logics; perhaps that future work may be able to exploit the progress of situation semantics in this area [Pólos 1993a, Pólos 1993b].

Bibliography

[Barwise&Perry 1983] Barwise, J. and Perry, J., *Situation and Attitudes*, (MIT Press, 1983).

[Biddle 1979] Biddle, B., *Role Theory*, (Academic Press, 1979).

[Blumer 1969] Blumer, H., *Symbolic Interactionism: Perspective and Methods*, (Englewood Cliffs, NJ, Prentice-Hall, 1969).

[Bond&Gasser 1988] Bond, A. and Gasser, L.,(eds.), *Readings in Distributed Artificial Intelligence*, (Morgan-Kaufmann, San Mateo, CA, 1988).

[Brown et al. 1991] Brown, A., Mantha, S. and Wakayama, T., Preferences as normative knowledge: towards declarative obligations, in: J-J. Ch. Meyer, R.J. Wieringa, (eds.), *Proceedings of DEON'91*, Free University of Amsterdam, (1991), 142-163.

[Carley 1986] Carley, K., Efficiency in a garbage can: implications for crisis management, in: J.G., March and R. Weissinger-Baylon, (eds.), *Ambiguity and Command*, (Marshfield, MA, Pitman, 1986), 165-194.

[Chisholm&Sosa 1966a] Chisholm, R., and Sosa, E., On the logic of "intrinsically better", *American Philosophical Quarterly* **3** (1966), 244-249.

[Chisholm&Sosa 1966b] Chisholm, R., and Sosa, E., Intrinsic preferability and the problem of supererogation, *Synthese* **16** (1966), 321-331.

[Cohen&Levesque 1987] Cohen, P., and Levesque, H., Persistence, intention and commitment, In: M. P. Georgeff and A. L. Lansky, (eds.), *Proceedings of the 1986 workshop on Reasoning about Actions and Plans*, (Morgan Laufmann Publishers, San Mateo, CA, 1987), 297-340.

[Cohen&Levesque 1990] Cohen, P., and Levesque, H., Intention is choice with commitment. *Artificial Intelligence* **42** (3) (1990).

[Danielsson 1968]
Danielsson, S., *Preference and Obligation*, (Filosofischa föreningen, Uppsala, 1968).

[Doyle 1983] Doyle, J., A Society of mind - multiple perspectives, reasoned assumptions, and virtual copies, *Proceedings of the 8th Inter. Joint Conference on AI*, Vol.I, (1983), 309-314.

[Doyle 1991] Doyle, J., Rational control of reasoning in artificial intelligence, in: A Fuhrmann, M. Morreau, (eds.), *The Logic of Theory Change*, (Springer Verlag, 1991), 19-48.

[Fagin&Halpern 1988] Fagin, R., and Halpern, J., Belief, awareness, and limited Reasoning, *Artificial Intelligence* **34** (1988) 39-76.

[Fagin&Vardi 1986] Fagin, R., and Vardi, M., Knowledge and implicit knowledge in a distributed environment: preliminary report, in: J. Y. Halpern, (ed.), *Proceedings of the First Conference on Theoretical Aspects of Reasoning about Knowledge*, (Morgan-Kaufmann, Los Altos, CA, 1986), 187-206.

[Fine&Schurz 1992] Fine, K., and Schurz, G., Transfer theorems for multimodal logics, in: *Proceedings of Arthur Prior Memorial Conference, Christchurch, New Zealand*, (to appear).

[French 1988] French, S., *Decision Theory, an Introduction to the Mathematics of Rationality*, (Ellis Horwood Limited, 1988).

[Gabbay 1992] Gabbay, D., *LDS - Labelled Deductive Systems*, (7th Draft), 1992.

[Gamut 1990] Gamut, L., *Logic, Language, and Meaning*, (The University of Chicago Press, 1990).

[Gärdenfors 1988] Gärdenfors, P., *Knowledge in Flux–Modeling the Dynamics of Epistemic States*, (MIT Press, Cambridge, Mass., 1988).

[Gärdenfors 1990] Gärdenfors, P., The dynamics of belief systems: foundations vs. coherence theories, *Revue Internationale de Philosophie* **1** 1990, 24-46.

[Gärdenfors&Makinson 1988] Gärdenfors, P., and Makinson, D., Revision of knowledge systems using epistemic entrenchment. In: M. Vardi, (ed.), *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, (Morgan Kaufmann, Los Altos, CA., 1988).

[Giddens 1979] Giddens, A., *Central Problems in Social Theory: Action, Structures, and Contradiction in Social Analysis*, (Berkeley, CA, University of California Press, 1979).

[Ginsberg 1986] Ginsberg, M., Counterfactuals, *Artificial intelligence* **30** (1986), 35-79.

[Ginsberg&Smith 1987] Ginsberg, M., and Smith, D., Reasoning about action I: a possible worlds approach, in: M. Ginsberg, ed., *Readings in Nonmonotonic Reasoning*, (Morgan Kaufmann, Los Altos, 1987).

[Grahne 1991] Grahne, G., Updates and counterfactuals, in: J. Allen, R. Fikes, and E. Sandewall, (eds.), *Proccedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, (Morgan Kaufmann Publishers, San Mateo, CA, 1991), 269-276.

[Habermas 1976] Habermas, J., *The Theory of Communicative Action*, (Beacon Press, Boston, 1984).

[Halldén 1957] Halldén, S., *On the Logic of Better*, (Library of Theoria 2, Lund, 1957).

[Halldén 1966] Halldén, S., Preference logic and theory choice, *Synthese* **16** (1966), 307-320.

[Halldén 1980] Halldén, S., *The Foundations of Decision Logic*, (CWK Gleerup,

Lund, 1980).

[Halpern&Fagin 1989] Halpern, J., and Fagin, R., Modelling knowledge and action in distributed systems, *Distributed Computing* **3** (1989), 159-177.

[Halpern&Moses 1992] Halpern, J., and Moses, Y., A guide to completeness and complexity for modal logics of knowledge and belief, *Artificial Intelligence* **54** (1992), 319-379.

[Hansson 1968] Hansson, B., Fundamental axioms for preference relations, *Synthese* **18** (1968), 423-442.

[Hansson 1989] Hansson, S., A new semantical approach to the logic of preference, *Erkenntnis* **31** (1989), 1-42.

[Harel 1984] Harel, D., Dynamic logic, in: D. Gabbay and F. Guenthner, (eds.), *Handbook of Philosophical Logic*, Vol.II, (D. Reidel Publishing company, 1984), 497-604.

[Hintikka 1962] Hintikka, J.,*Knowledge and Belief*, (Cornell University Press, 1962).

[Hirofumi&Mendelzon 1991] Hirofumi, K., and Mendelzon, A., On the difference between updating a knowledge base and revising it, in: J. Allen, R. Fikes, and E. Sandewall, (eds.), *Proccedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, (Morgan Kaufmann Publishers, San Mateo, CA, 1991), 387-394.

[van der Hoek et al., 1993] van der Hoek, W., van Linder, B., and Meyer, J.-J. Ch., A logic of capabilities, Free University Amsterdam, Report IR-330, 1993.

[Huang 1989] Huang, Z., Dependency of belief in distributed systems, in: M. Stokhof and L. Torenvliet, (eds.) *Proceedings of the Seventh Amsterdam Colloquium*, (1990), 637-662.

[Huang&Kwast 1990] Huang, Z., and Kwast, K., Awareness, negation and logical omniscience, *Lecture Notes in Artificial Intelligence* **478** (1991), 282-300.

[Huang 1990] Huang, Z., Logics for belief dependence, *Lecture Notes in Computer Science* **533** (1991), 274-288.

[Huang 1990a] Huang, Z., General epistemic logic and its problem, *Philosophical Research* **5** (1990) 91-93. (in Chinese). Also available in: CCSOM Reprint 90-04.

[Huang&van Emde Boas 1990] Huang, Z., and van Emde Boas, P., Belief dependence, revision and persistence, in: *Proceedings of the Eight Amsterdam Colloquium*, (1992), 253-270.

[Huang&van Emde Boas 1991] Huang, Z., and van Emde Boas, P., Schoenmakers paradox: its solution in a belief dependence framework, Institute for Language, Logic and Computation, University of Amsterdam, Preprint LP-91-05.

[Huang 1991a] Huang, Z., Reasoning about knowledge, *Computer Science* **1** (1991) 46-48. (in Chinese). Also available in: CCSOM Reprint 91-22.

[Huang 1991b] Huang, Z., Theory and methods of belief revision, *Computer Science* **6** (1991) 12-16. (in Chinese). Also available in: CCSOM Reprint 91-48.

[Huang&Masuch 1991] Huang, Z., and Masuch, M., Reasoning about action: a comparative survey, CCSOM Research Report 91-37, (1991).

[Huang, Masuch&Pólos 1992a] Huang, Z., Masuch, M., and Pólos, L., A preference

logic for rational actions, in: R. Blanning and D. King, (eds.), *Artificial Intelligence in Organization Design, Modeling and Control*, Information Systems Series, IEEE Computer Society Press (forthcoming).

[Huang, Masuch&Pólos 1992b] Huang, Z., Masuch, M., and Pólos, L., Een preferentie-logica voor rationele handelingen, in: Swaan Arons, H. de, H. Koppelaar and E. J. H. Kerckhoffs, (eds.). *Conferentie Proceedings of NAIC '92*, (1992), 17-28.

[Huang, Masuch&Pólos 1992c] Huang, Z., Masuch, M., and Pólos, L., ALX: an action logic for agents with bounded rationality, *Artificial Intelligence* (forthcoming).

[Huang 1992] Huang, Z., Autoepistemic logic and non-monotonic reasoning, *Computer Science* **4** (1992) 9-13. (in Chinese). Also available in: CCSOM Reprint 92-69.

[Huang, Masuch&Pólos 1993] Huang, Z., Masuch, M., and Pólos, L., ALX2: the quantifier ALX logic, CCSOM Research Report 93-99, (1993).

[Huang&Masuch 1993a] Huang, Z., and Masuch, M., ALX3: a multi-agent ALX logic, in: Müller, A., and H. W. M. Gazendam (eds.), *Multi-agent Systems*, (Berlin and New York: De Gruyter) (forthcoming).

[Huang&Masuch 1993b] Huang, Z., and Masuch, M., Reasoning about action, *Computer Science* **3** (1993) 7-13. (in Chinese). Also available in: CCSOM Reprint 93-103.

[Huang&van Emde Boas 1994] Huang, Z., and van Emde Boas, P., Information acquisition from multi-agents resources, in: *Proceedings of the 5th Conference on Theoretical Aspects of Reasoning about Knowledge*, (Morgan Kaufmann, Los Altos, CA., 1994), 65-79.

[Huang 1994] Huang, Z., New advances in reasoning about knowledge, *Computer Science* **3** (1994), 49-52. (in Chinese).

[Hughes 1984] Hughes, G., and Cresswell, M., *A Companion to Modal Logic*, (Methuen, London and New York, 1984).

[Jackson 1972] Jackson, J., *Role.* (Cambridge University Press, 1972).

[Jackson 1991] Jackson, F.,(ed.), *Conditionals*, (Oxford University Press, 1991).

[Jackson 1989] Jackson, P., On the semantics of counterfactuals, *Proceedings of IJCAI-89*, (Detriot, Michigan, USA, 1989).

[Jeffrey 1983] Jeffrey, R., *The Logic of Decision*, (2nd edition), (New York, 1983).

[Konolige 1983] Konolige, K., A deductive model of belief, *Proceedings of the 8th International Joint Conference on AI*, (1983), 377-381.

[Konolige 1986] Konolige, K., What awareness isn't: a sentential view of implicit and explicit belief, in: J. Halpern (ed.) *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the 1986 Conference*, (Morgan-Kaufmann, Los Altos, CA, 1986), 241-250.

[Kuhn 1974] Kuhn, A., *The Logic of Social System*, (Jossey-Bass, London, 1974).

[Kwast 1992] Kwast, K., *Unknown Values in Retational Database Systems*, Ph.D. thesis, University of Amsterdam, (1992).

[Levesque 1984] Levesque, H., A logic of implicit and explicit belief, in: *Proceedings*

AAAI-84, Austin, TX (1984), 198-202.

[Lewis 1973] Lewis, D., *Counterfactuals*, (Blackwell, Oxford, 1973).

[Luhmann 1982] Luhmann, N., *The Differentiation of Society*, (New York, Columbia University Press, 1982).

[March 1976] March, J., The technology of foolishness, In: J. G. March and J. P. Olsen, (eds.), *Ambiguity and Choice in Organizations*, (Bergen, Norway, Universitetsforlaget, 1976), 69-81.

[March&Olsen 1986] March, J., and Olsen, J., Garbage can models of decision making in organizations, in: J. G. March and R. Weissinger-Baylon, (eds.) *Ambiguity and Command*, (Marshfield, MA, Pitman, 1986), 11-53.

[Martin&Shapiro 1986] Martin, J., and Shapiro, C., Theoretical foundations for belief revision, in: J. Y. Halpern, (ed.) *Proceedings of the First Conference on Theoretical Aspects of Reasoning about Knowledge* (Morgan-Kaufmann, Los Altos, CA, 1986), 383-398.

[Marx et al., 1993] Marx, M., Huang, Z., and Masuch, M., A new preference logic, CCSOM Research Report 92-73, (1993).

[McCarthy&Hayes 1969] McCarthy, J. and Hayes, P., Some philosophical problems from the standpoint of AI, in: B. Meltzer and D. Michie, (eds.), Machine Intelligence, Vol. IV, (Edinburgh University Press, 1969).

[Masuch 1991] Masuch, M., Formalization of Thompson's Organization in Action, CCSOM Research Report 91-32, (1991).

[Masuch&Huang 1994] Masuch, M. and Huang, Z., A logical deconstruction of organizational action: formalizing J. D. Thompson's *Organization in Action* in a multi-agent action logic, CCSOM Working Paper 94-120.

[Meyer 1989] Meyer, J.-J. Ch., An analysis of the Yale shooting problem by means of dynamic epistemic logic, Free University Amsterdam, report IR-201, 1989.

[Moore 1985] Moore, R., A formal theory of knowledge and action, report No. CSLI-85-31, 1985.

[Mullen 1979] Mullen, J., Does the logic of preference rest on a mistake, *Metaphilosophy* **10** (1979), 247-255.

[Nebel 1990] Nebel, B., *Reasoning and Revision in Hybrid Representation Systems*, Lecture Notes in Computer Science **422** (1990).

[Nute 1986] Nute, D., Conditional logic, in: D. Gabby and F. Guenthner, (eds.), *Handbook of Philosophical Logic*, Vol.II, (1986), 387-439.

[Padgett 1980] Padgett, J., Managing garbage can hierarchies, *Administrative Science Quarterly* **25** (1980), 583-604.

[Parsons 1937] Parsons, T., *The Structure of Social Action*, (Glencoe, IL, Free Press, 1937).

[Pearl 1988] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, (Morgan Kaufmann Publishers, San Mateo, CA, 1988).

[Pollock 1976] Pollock, J., *Subjunctive Reasoning*, (Reidel, Dordrecht, 1976).

[Pollock 1981] Pollock, J., A refined theory of counterfactuals, *Journal of Philosophical Logic* **10** (1981), 239-266.

[Pólos 1993a] Pólos, L., Information states in situation semantics, CCSOM Research

Report 93-98, (1993).

[Pólos 1993b] Pólos, L., Updated situation semantics, *Journal of Symbolic Logics*, (forthcoming).

[Pólos et al., 1993] Pólos, L., Masuch, M., and Huang, Z., A hierarchical semantics for the logic of preferences, CCSOM Research Report 93-98, (1993).

[Pörn 1989] Pörn, I., On the nature of social order, in: Fenstad, et al., (eds.) *Logic, Methodology, and Philosophy of Science, VIII*, (Elsviere, 1989).

[Potts et al., 1989] Potts, G., John, M., and Kirson, D., Incorporating new information into existing world knowledge, *Cognitive Psychology* **21** (1989), 303-333.

[Rao&Georgeff 1991] Rao, A., and Georgeff, M., Modeling rational agents within a BDI- architecture, in: J. Allen, R. Fikes, and E. Sandewall, (eds.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, (Morgan Kaufmann Publishers, San Mateo, CA, 1991), 473-484.

[Rescher 1967] Rescher, N., Semantic foundations for the logic of preference, in: N. Rescher, (ed.), *The Logic of Decision and Action*, (University of Pitteburgh press, 1967).

[Rescher 1967a] Rescher, N., (ed.), *The Logic of Decision and Action*, (University of Pitteburgh Press, 1967).

[Schutz 1967] Schutz, A., *The Phenomenology of the Social World*, (Evanston, IL, Northwestern University Press, 1967).

[Schoenmakers 1986] Schoenmakers, W., A problem in knowledge acquisition, *SIGART Newsletter* **95** (1986), 56-57.

[Scott 1970] Scott, D., Towards a mathematical theory of computation, *Proceedings of 4th Annual Princeton Conference on Information Science and Systems*, (1970) 169-176.

[Scott 1982] Scott, D., Domains for denotational semantics, in: M. Nielsen and E. T. Schmidt, (eds.), *Lecture Notes in Computer Science* **140** (1982), 577-613.

[Simon 1955] Simon, H., A behavioral model of rational choice, *Quarterly Journal of Economics* **69** (1955), 99-118.

[Simon 1964] Simon, H., On the concept of organizational goal, *ASQ* **9** (1964), 1-22.

[Stalnaker 1968] Stalnaker, R., A theory of conditionals, in: *Studies in Logical Theory, American Philosophical Quarterly* **2** (1968), 98-122.

[Thijsse 1992] Thijsse, E., *Partial Logic and Knowledge Representation*, Ph. D. thesis, 1992.

[Thompson 1967] Thompson, J., *Organizations in Action, Social Science Bases of Administrative Theory*, (Mc Graw-hill Book Company, 1967).

[Werner 1988] Werner, E., Toward a theory of communication and cooperation for multiagent planning, in: M. Y. Vardi, (ed.) *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge,* (Morgan Kaufmann, Los Altos, CA, 1988), 129-143.

[Werner 1990] Werner, E., Cooperating agents: a unified theory of communication and social structure, in: L. Gasser and M. H. Huhns, (eds.) *Distributed Artificial Intelligence: Vol. II*, (Morgan Kaufmann Publishers, 1990), 4-37.

[Winslett 1988] Winslett, M., Reasoning about action using a possible models approach. In: *Proceedings of AAAI-88*, (ST. Paul, Minnesota, 1988), 89-93.

[von Wright 1963] von Wright, G., *The Logic of Preference*, (Edinburgh, 1963).

[von Wright 1972] von Wright, G., The Logic of Preference Reconsidered, *Theory and Decision* **3** (1972), 140-169.

Index

List of symbols

Sets, functions and relations

Samenvatting

Begrensde rationaliteit (*bounded rationality*) heeft twee interpretaties: een ruime en een beperkte. In de ruime interpretatie refereert begrensde rationaliteit aan het fenomeen dat actoren (*agents*) een beperkte hoeveelheid cognitieve middelen en vaardigheden hebben. In de beperkte interpretatie refereert begrensde rationaliteit aan de term die door H. A. Simon geïntroduceerd is. Hij gaat uit van een beslisprocedure voor rationele actoren die niet alle mogelijke handelingen kennen, noch hun precieze gevolgen, en niet beschikken over een volledige preferentie ordening voor deze gevolgen. In dit proefschift zal ik verschillende logicas voor actoren met een begrensde rationaliteit presenteren en bespreken.

Metbetrekking tot de ruime interpretatie van begrensde rationaliteit zal ik mij concentreren op het fenomeen van geloofsafhankelijkheid (*belief dependence*) in een meedere-actoren omgeving, hierbij refeert geloofsafhankelijkheid aan het fenomeen dat bepaalde actoren afhankelijk zijn van andere met betrekking tot hun geloof, kennis, of informatie, als gevolg van hun eigen beperkingen.

De beperkte interpretatie van begrensde rationaliteit zal ik toespitsen op de bestudering van handelingslogicas (*action logics*) voor actoren met H. A. Simons begrensde rationaliteit. Deze studie mondt uit in de ontwikkeling van een formele taal voor de sociale wetenschappen, met name voor organizatietheorieën.

Het proefschrift is verdeeld in twee delen: Deel I "*Logics for Belief Dependence*" en Deel II "*Action Logics for Agents with Bounded Rationality*". Deel I bestudeert logicas voor geloofsafhankelijkheid, en een studie van begrensde rationaliteit in de ruime interpretatie. Deel II bestudeert handelingslogicas voor actoren met begrensde rationaliteit in de beperkte interpretatie.

In Deel I behandelt Hoofdstuk 2 de termen rationaliteit en begrensde rationaliteit, onderzoekt de varianten van begrensde rationaliteit, en bespreekt het belang van hun toepassing. Een syntax voor geloofsafhankelijkheidslogicas wordt ingeroerd, en een algemeen scenario voor het formalizeren van geloofsafhankelijkheid wordt gepresenteerd. In Hoofdstuk 3, worden meerdere logicas voor geloofsafhankelijkheden voorgesteld. Ik laat zien dat de voorgestelde logicas voldoende uitdrukkingkracht hebben voor de formalizering van communicatie tussen meerdere actoren met beperkte

informatie. Hoofdstuk 4 introduceert de semantische modellen voor deze logicas. Ik beweer dat algemene epistemische en doxatische logicas niet geschikt zijn om met geloofsafhankelijkheid om te gaan. Vervolgens stel ik meerdere semantische modellen voor, vergelijk deze modellen, en bespreek onder welke omstandigheden deze modellen kunnen worden toegepast. In Hoofdstuk 5 bestudeer ik de geloofsdynamica in het raamwerk van geloofsafhankelijkheid. Gebruik makend van de geloofsafhankelijkheidslogicas stel ik een mechanisme voor om te berekenen hoe een actor een keuze kan maken uit verschillende geloofsherzieningsalternatieven, zoals geloofswijziging, -uitbreiding, verwerving, en -behoud. In Hoofdstuk 6 bestudeer ik een probleem dat door W. Schoenmaker geïntroduceerd is, en dat karakteristiek is voor de bestudering van informatievergaring uit meerdere bronnes. Een algemene benadering voor het formalizeren van het probleem van informatieverwerving uit meerdere bronnen wordt gepresenteerd. Een aantal begrippen met betrekking tot Schoenmakers' probleem, zoals "absolute-zekerheid", "zekerheid", en "sterke-zekerheid", worden formeel gedefinieerd. Bovendien, gebruikmakend van de logica voor geloofsafhankelijkheid, wordt er een begrip voor "bijna-zekerheid" gedefinieerd, dat een redelijke en acceptable strategie voor het Schoenmakers probleem oplevert. In Hoofdstuk 7 worden uitbreidingen op de geloofsafhankelijkheidslogicas besproken, en worden de conclusies van Deel I geformuleerd.

Deel II begint met Hoofdstuk 8 waarin ik algemene ideeën over handelingslogicas voor actoren met begrensde rationaliteit in het kader van een formele taal voor de sociale wetenschappen presenteer. In Hoofdstuk 9 wordt de term "preferentie" onderzocht, en wordt onderscheid gemaakt tussen vier vormen van preferentie relaties, te weten "actuele-preferentie", "ceteris-paribus-preferentie", "conditionele-preferentie", en "absolute-preferentie". Bovendien geef ik zowel syntactische als semantische kenmerken van deze preferentie relaties. Tenslotte worden de corresponderende logicas beschouwd. In Hoofdstuk 10 wordt een eenvoudig systeem voorgesteld, dat preferentie logica, herzieningslogica (*update logics*), en propositionele dynamische logica combineert. De gezondheid en volledigheid van dit systeem wordt bewezen. Bovendien worden handelingen die aanleiding geven tot minimale veranderingen bestudeerd. Hoofdstuk 11 behandelt een multi-modale predicaat-logische versie van ALX logica. In dit hoofdstuk presenteer ik enige aspecten van het toepassen van meerdere-actoren handelingslogicas. Ik bospreek een aantal voor de hand liggende toepassingen van ALX voor een formele theorie van sociale actoren (*social agents*), en laat zien dat de ALX3 logica inderdaad genoeg uidrukkingskracht heeft voor zinvolle applicaties.

Titles in the ILLC Dissertation Series:

Transsentential Meditations; Ups and downs in dynamic semantics
Paul Dekker
ILLC Dissertation series, 1993-1

Resource Bounded Reductions
Harry Buhrman
ILLC Dissertation series, 1993-2

Efficient Metamathematics
Rineke Verbrugge
ILLC Dissertation series, 1993-3

Extending Modal Logic
Maarten de Rijke
ILLC Dissertation series, 1993-4

Studied Flexibility
Herman Hendriks
ILLC Dissertation series, 1993-5

Aspects of Algorithms and Complexity
John Tromp
ILLC Dissertation series, 1993-6

The Noble Art of Linear Decorating
Harold Schellinx
ILLC Dissertation series, 1994-1

Generating Uniform User-Interfaces for Interactive Programming Environments
Jan Willem Cornelis Koorn
ILLC Dissertation series, 1994-2

Process Theory and Equation Solving
Nicoline Johanna Drost
ILLC Dissertation series, 1994-3

Calculi for Constructive Communication, a Study of the Dynamics of Partial States
Jan Jaspars
ILLC Dissertation series, 1994-4

Executable Language Definitions, Case Studies and Origin Tracking Techniques
Arie van Deursen
ILLC Dissertation series, 1994-5

Chapters on Bounded Arithmetic & on Provability Logic
Domenico Zambella
ILLC Dissertation series, 1994-6

Adventures in Diagonalizable Algebras
V. Yu. Shavrukov
ILLC Dissertation series, 1994-7

Learnable Classes of Categorial Grammars
Makoto Kanazawa
ILLC Dissertation series, 1994-8

Clocks, Trees and Stars in Process Theory
Wan Fokkink
ILLC Dissertation series, 1994-9

Logics for Agents with Bounded Rationality
Zhisheng Huang
ILLC Dissertation series, 1994-10

On Modular Algebraic Prototol Specification
Jacob Brunekreef
ILLC Dissertation series, 1995-1