FORMAL MODELS OF DIALOGUE PARTICIPANTS

対話行為者の形式的モデルに関する研究

杉本 徹

FORMAL MODELS OF DIALOGUE PARTICIPANTS

①

FORMAL MODELS OF DIALOGUE PARTICIPANTS

対話行為者の形式的モデルに関する研究

by

Toru Sugimoto

杉本 徹

Doctoral Thesis

Submitted to

the Graduate School of
the University of Tokyo
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Science
in Information Science

January 1995

Abstract

As a general basis for constructing a flexible and cooperative natural language dialogue system, we need a model of dialogue participants, in which knowledge for various dialogue tasks is represented in a common declarative form and a general inference mechanism manipulates representation of mental states. We present two such models which meet two important requisites for representation systems. The first model represents an agent's mental states in the form called Mental World Structure, which has strong expressive power for modalities. We can deal with composition, quantification and unification of modalities, and thus we can concisely express various types of knowledge that are difficult to express in previous representation systems. We smoothly incorporate into our framework three basic inference procedures, that is, deduction, abduction and truth maintenance. With them, we explain an agent's inference processes working behind cooperative dialogues. Inferences about plans are modeled clearly, using modalities of belief and time together. The second model is a logic of mental attitudes based on preference ordering. An agent's model structure includes explicit representations of preferences, that is, the plausibility order and the desirability order. Mental attitudes such as belief, intention and choice are defined in terms of the preference orders. In particular, defined intentions satisfy most requisites ever proposed such as freedom from consequential closure and persistency. We introduce operators for preference between sentences, and examine their properties and relationship with other attitudes. Then we apply this logic to reasoning about plans. We give a formal account of plan construction and selection processes, and we examine several heuristics for plan recognition currently used.

Acknowledgements

I would like to express my appreciation to my advisor, Professor Akinori Yonezawa. He has given me many valuable suggestions and encouraged me all the time. I am grateful to Takeshi Fuchi, Takashi Miyata, and Kentaro Torisawa for many fruitful discussions on natural language understanding. I am also grateful to the members of our laboratory. They have been so helpful and encouraged me.

For Part 1 of the thesis, I would like to thank Ichiro Ohsawa and Hideyuki Nakashima, for giving me useful comments on earlier versions of it. This research is based on my earlier research at Tokyo Institute of Technology. I would like to thank Professor Masako Horai Takahashi and the members of Horai laboratory for giving me many valuable comments on my research and discussions on related topics.

For Part 2 of the thesis, I would like to thank members of Cognitive Science Group at Electrotechnical Laboratory, who gave me many useful comments on this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

As a general basis for constructing a flexible and cooperative natural language dialogue system, we need to model dialogue participants: their reasoning, mental states such as beliefs and intentions, together with their dynamic change. A dialogue participant performs various tasks. First, he understands his dialogue companion's utterances. He is flexible to interpret possibly ambiguous linguistic expressions using contexts of the utterances. Second, he recognizes his companion's intentions and plans behind the utterances. By extracting information conveyed indirectly, he can behave cooperatively toward his companion. Third, he forms intentions and constructs his own plans. A dialogue participant usually constructs plans of making an utterance in order to inform or request his companion something. Fourth, he generates linguistic expressions and utters them. He chooses appropriate expressions based on his knowledge about his companion and contexts.

Traditionally, these tasks are processed by separate modules using specialized representation frameworks and specialized inference mechanisms. Semantic interpretation models [19, 30, 38] use specialized frameworks for representing semantic information and inference procedures to choose plausible interpretations and referents. Plan recognition models [1, 6, 21] and plan construction systems [3, 4, 15] use specialized frameworks for representing properties of actions and have procedures that find preferable plans. Linguistic generation models [4, 22] use specialized grammar and specific inference procedures to generate helpful and clear responses. Behind these studies exists a typical model of dialogue participants, which we call a *traditional* model of dialogue participants (see Figure 1.1). It consists of four (or more) separate modules connected serially. When a dialogue participant gets an input (that is, recognizes an utterance), he first uses the semantic interpretation module, then uses the plan recognition module, the plan construction module, and finally uses the linguistic generation module to output a response.

A traditional model of dialogue participants has the following problems:

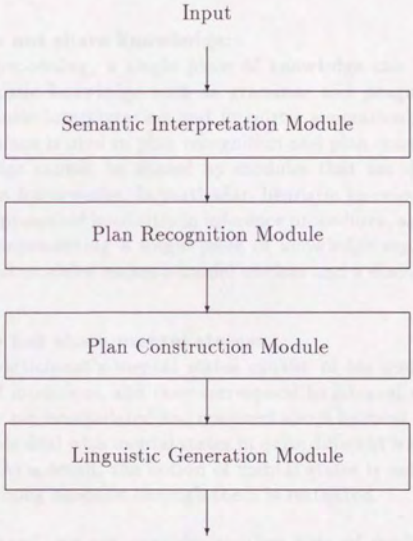


Figure 1.1: A Traditional Model of a Dialogue Participant

1. **Modules work only in a fixed order and interact in a fixed way:**
   Since modules are designed independently with various assumptions, they work only in a fixed order and interact in a fixed way, even if possible. This obstructs flexibility of dialogue participants' models. Dialogue tasks need to interact with each other during processing. For example, the semantic interpretation task often requires information about the speaker's plans obtained in later plan recognition task. Furthermore, when semantic interpretation fails, the control should go directly to linguistic generation, that is, generation of a clarifying response.

2. **Modules do not share knowledge:**
   In dialogue processing, a single piece of knowledge can be used in several tasks. Linguistic knowledge such as grammar and pragmatic convention is used in semantic interpretation and linguistic generation. Knowledge about actions and plans is used in plan recognition and plan construction. However, such knowledge cannot be shared by modules that use different specialized representation frameworks. In particular, heuristic knowledge and preferences are usually represented implicitly in inference procedures, and thus they cannot be shared. Representing a single piece of knowledge separately in different form in several modules makes a model unclear and a dialogue system hard to maintain.

3. **Modules do not share mental states:**
   A dialogue participant's mental states consist of his mental attitudes such as beliefs and intentions, and they correspond to internal states of a dialogue system. They are manipulated and reasoned about by most modules. However, existing models deal with mental states in quite different ways and they cannot share them. As a result, the notion of mental states is not used enough, and interaction among modules through them is restricted.

On the other hand, we can consider another type of model of dialogue participants, which we investigate in this thesis (see Figure 1.2). Dialogue tasks are regarded as a kind of problem solving and processed together in a general formal framework for problem solving. Knowledge for these tasks is represented in a common declarative form and used by a general inference mechanism. This model is formal in the sense that represented knowledge has clear (formally defined or intuitive) semantics and is used in accordance with it. Mental states are also represented declaratively, and they are referred to and manipulated by the inference mechanism. We think that this approach will give us a flexible and clear model of a dialogue participant. Dialogue tasks can be processed in arbitrary order and interact freely with each other. A single piece of knowledge and a mental state can be used in various tasks.

Providing a representation framework that is expressive enough to represent knowledge for dialogue processing is particularly important in this approach. How-

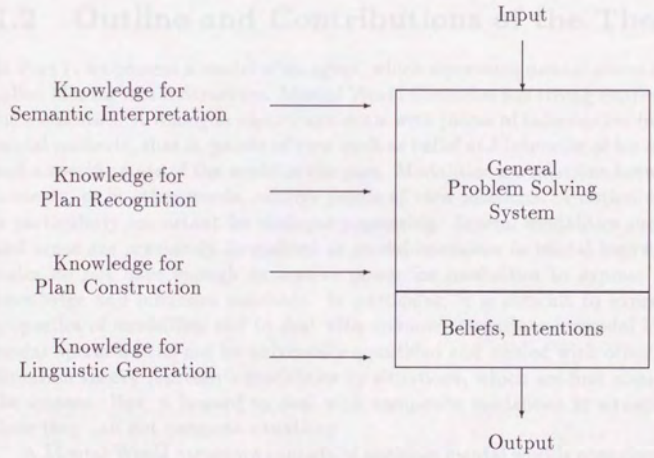


Figure 1.2: An Integrated Model of a Dialogue Participant

ever, existing models such as Hasida's dependency propagation [18], Hobbs's abductive inference schema [20] and Konolige's direct argumentation system [24] simply use standard first-order logic, and requisites for such a representation framework are not fully studied yet. We think there are two important requisites:

1. A representation framework needs strong expressive power for modalities.

2. A representation framework can deal with preferences of an agent.

In this thesis, we present two formal models of dialogue participants, which meet these requisites.

1.2 Outline and Contributions of the Thesis

In Part 1, we present a model of an agent, which represents mental states in the form called Mental World Structure. Mental World Structure has strong expressive power for modalities. A dialogue participant deals with pieces of information from various modal contexts, that is, points of view such as belief and intention of his companion, and a specific state of the world in the past. Modalities are relations between modal contexts, or in other words, relative points of view relations. A notion of modality is particularly important for dialogue processing. Several modalities such as belief and tense are previously formalized as modal operators in modal logics, but these logics do not have enough expressive power for modalities to express an agent's knowledge and inference concisely. In particular, it is difficult to express general properties of modalities and to deal with unknown situations in modal logics, since modal operators can not be universally quantified and unified with other operators. Situation theory represents modalities by situations, which are first class objects of the system. But, it is hard to deal with composite modalities in situation theory, since they can not compose situations.

A Mental World Structure consists of multiple mental worlds organized in a tree. Each mental world is a set of mental propositions and corresponds to one modal context. Modalities are represented by path expressions, which are first class syntactic objects. Path expressions are used in one mental world to refer to another world. We can manipulate path expressions in various ways, and we show that we can easily express phenomena that are difficult to deal with in previous representation systems. First, we can compose path expressions to make up one composite path expressions. Thus, composite modalities and simple ones are treated in a similar way, and description of knowledge is simplified. Second, using path expression variables, we can deal with quantification over modalities. We can express commonsenses, which hold in all contexts. Third, we can unify path expressions. For example, we represent an unknown situation such as the reference time of an utterance by a newly created path expression, and unify it with another one afterward.

The inference mechanism applies equally to each mental world, and it adds or deletes propositions. We adopt three procedures as the basic inference procedures,

that is, deduction, abduction, and truth maintenance. An agent's all inference processes are modeled by successive applications of these three procedures in an appropriate order. The deduction procedure is a sound inference procedure that uses implications forwardly to obtain new information. On the other hand, the abduction procedure uses implications backwardly to find an explanation of observed facts. A dialogue participant does not always draw sound inference, and thus nondeterministic procedures like the abduction procedure are indispensable. The truth maintenance procedure is used to recover consistency of a Mental World Structure when it becomes inconsistent for some reason.

Then, we explain an agent's inference process working behind cooperative dialogues in our framework. We show that we can deal with inference about plans in a natural way by regarding intentions as a kind of beliefs about the future. The problem of inference control, that is, the problem of determining application order of inference procedures and selecting the best explanation of a given fact is in general very hard to solve. We discuss it in some detail and explain the need of various kinds of preferences about the domain.

The main contributions of Part 1 are as follows:

1. We present a new formal framework for representing mental states of an agent called Mental World Structure, which has strong expressive power for modalities. Modalities can be composed, quantified and unified, and various types of knowledge that are difficult to express in previous representation systems are shown to be expressed concisely in our framework.

2. We smoothly incorporate into our framework three basic inference procedures, that is, deduction, abduction and truth maintenance. They are defined relative to mental worlds, hence are applied equally to each mental world.

3. We demonstrate the strength of our representation system and inference procedures by providing an explanation of an agent's inference processes working behind example cooperative dialogues. Only a small set of clear background beliefs is used.

In Part 2, we present a model of an agent in the form of a logic of mental attitudes based on preference ordering. Preferences are evaluations about plausibility and desirability of propositions or worlds. A dialogue participant uses various preferences to understand and generate linguistic expressions and to recognize plans of his companion. In the traditional model of dialogue participants, these preferences are represented implicitly in inference procedures. However, to deal with sharing and interactions of preferences, we need a general explicit representation framework for them.

In this thesis, we deal with qualitative preferences, which are explicitly represented by partial orders on model structures. An agent's mental state is specified by knowledge and two preference orders, that is, the plausibility order and the desirability order. The language of our logic is a propositional language extended

by adding attitudinal operators: belief, intention, choice, and preference between sentences. The satisfaction relation for these operators is defined in terms of the preference orders. Besides mental attitudes about the states of the world, mental attitudes about another agent's mental states can be dealt with. Furthermore, we introduce a construct of sentences, which is used to specify an agent's knowledge and the preference orders. Then we apply this logic to reasoning about plans. We give a formal account of plan construction and selection process, and examine several heuristics for plan recognition currently used.

The main contributions of Part 2 are as follows:

1. We present a model of an agent based on two preference orders: one is about plausibility and the other is about desirability. Mental attitudes such as belief, intention and choice are defined in terms of preference orders.

2. We introduce operators for preference between sentences, and examine their properties. We give several types of preferences frequently used, and show that the strongest type of them is closely related to other mental attitudes such as belief and intention.

3. We give an intuitive formulation of a notion of intention which satisfies most of requisites ever proposed such as freedom from consequential closure and persistency.

4. We give a formal account of plan construction and selection processes. We model plan construction with multiple preferences and dynamic revision of plans.

5. We examine several heuristics for plan recognition currently used. Giving preferences that validate widely used heuristics, we demonstrate that our framework can give a good formal basis for plan recognition models.

This thesis is structured as follows: Part 1 starts from Chapter 2 and ends in Chapter 9. In Chapter 2, we give an introduction and overview of Part 1. Chapter 3 introduces a notion of modalities in an agent's mental states and discusses problems with existing representation systems for modalities. Chapter 4 gives the formal definition of Mental World Structure, and Chapter 5 describes the three basic inference procedures. In Chapter 6, we present examples of cooperative dialogues, and analyze them in our framework. Chapter 7 compares our framework with situation oriented programming language PROSIT, and Chapter 8 discusses the problem of inference control. Chapter 9 concludes Part 1.

Part 2 starts from Chapter 10. In Chapter 10, we give an introduction of Part 2. Chapter 11 discusses need of preference in models of dialogue participants. In Chapter 12, we gives syntax and semantics of our logic and then examine properties of beliefs, choices and preferences. Chapter 13 deals with a notion of intention and a generalized version of it. In Chapter 14, we apply our logic to reasoning about

plans, that is, plan construction and plan recognition. Finally, Chapter 15 concludes Part 2. Proofs of main theorems in Part 2 are given in the appendix.

Part I

Multiple World Representation of Mental States for Dialogue Processing

Chapter 2

Introduction

In order to construct a cooperative and flexible dialogue system, we need to model powerful inference capability of an agent who participates in a dialogue. Such an agent draws various types of inference, that is, interprets possibly incomplete linguistic expressions, recognizes speaker's plans, and generates own appropriate goals and actions, among others, to communicate with other agents successfully. There are a number of independent models for each of these types of inference, but most of them such as semantic interpretation models [30] and plan-based dialogue processing models [1, 4] use specialized frameworks for knowledge representation and inference. This is a severe limitation for those models, since it is hard for them to realize interaction of different types of inference process which is essential for agent's inference capability.

In Part 1 of this thesis, we present a new formal unified framework for agent's problem solving, which can be used to explain his inference processes working behind various phenomena in dialogues. Like other logic-based problem solving models, our framework has two parts, one is declarative representation of knowledge or mental states, and the other is the inference mechanism.

As for representing agent's mental states, two points are important. First, simple first order language is not sufficient since as agent's mental objects, relations and propositions should be treated as first class citizens of the representation. Second, an agent usually has beliefs in various points of view, and strong expressive power for modalities is needed. Modalities should also be treated as first class objects, so ordinary formal modal logics are insufficient for this purpose. In our framework, we represent an agent's mental states in the form called *Mental World Structure*, or *MWS* for short, which consists of multiple *mental worlds*. Each mental world is viewed as a set of mental propositions and corresponds to one modal context, that is, specific point of view. With MWS, we can handle modalities more flexibly than ordinary modal logics and other representation systems.

Dialogue participants, or in general, agents in daily life draw not only sound deductive inference, but also possibly incorrect and creative inference, such as abductive one. Although many problem solving models only use backward-chained

deduction, some researchers try to model an agent's general inference ability naturally, to mention a few, Hasida's dependency propagation [18] and Hobbs's abductive inference schema [20] for natural language applications. Following these approaches, we adopt three distinct procedures as the basic inference procedures, that is, *deduction*, *abduction* and *truth maintenance*, and integrate them into our representation structure. We believe this set of inference procedures suffices for explaining an agent's inference processes in most of cooperative natural language dialogues.

Chapter 3

Modalities in Mental States

3.1 Modal Contexts

An agent usually has various pieces of information from many points of view. For instance, an agent may have pieces of information that belong to the following viewpoints:

(W_1) his own beliefs,
(W_2) his own beliefs about another agent's, say A's beliefs,
(W_3) his own beliefs about his own intentions,
(W_4) his own beliefs about a specific state of the world in the past, and
(W_5) his own beliefs about his own intentions about A's beliefs.

When information P belongs to the point of view W_1, the agent believes that P holds. Similarly when P belongs to W_2, he believes that A believes that P holds, and when P belongs to W_3, he believes that he intends to make P true, namely, he in fact intends to make P true. Each point of view corresponds to a possibly *partial* state of the world. For example, W_1 corresponds to the state of the world based on the agent's own beliefs about the current world, W_2 corresponds to the state based on his own beliefs about A's beliefs, and W_3 corresponds to the state into which he intends to change the current world state. Hence an agent has representations of many states of the world simultaneously in his mind. We call here each of these points of view or representations of states of the world a *modal context*. As representations of states of the world, modal contexts should always be consistent and closed under logical consequence.

It is often the case that a modal context corresponds to one point of view contained in another modal context. For instance, modal context W_2 corresponds to a point of view i.e., A's beliefs contained in W_1. With respect to this relative "point of view" relation, the set of modal contexts in the representation forms a hierarchy, as illustrated in Figure 3.1. Here each box represents a modal context and each arc represents a point of view relation between two contexts. Labels $bel(A)$, $int(i)$ and $past\_1$ represent kinds of the relation, that is, A's beliefs, agent's own intentions,

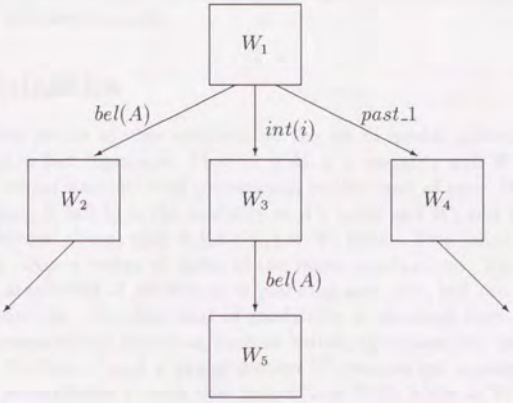


Figure 3.1: A Hierarchy of Modal Contexts

and a specific past instant of time, respectively. We call W_1, the top context of the hierarchy or the *base* context. It represents agent's own beliefs of the most simple type, and all other contexts are parts of this context.

3.2 Local Reasoning

Having representations of many states of the world simultaneously in his mind, an agent can reason about each state i.e., modal context alike. As an illustration of this fact, consider the following pair of simple deductive inferences drawn by an agent.

$$\frac{P \qquad P \supset Q}{Q} \qquad\qquad \frac{\text{I intend to make } A \text{ believe } P \qquad \text{I intend to make } A \text{ believe } P \supset Q}{\text{I intend to make } A \text{ believe } Q}$$

These two inferences are essentially the same processes, since they are both simple applications of the modus ponens rule. What differs is the context where inferences are done, that is, W_1 and W_5, respectively. So we think these two inferences should be done in the same manner with similar inferential costs, and call this type of context-relative reasoning *local reasoning*. In order to realize local reasoning, repre-

sentation frameworks for agent's mental states must have explicit representations of the hierarchy of modal contexts.

3.3 Modalities

We call relative points of view relations on the set of modal contexts *modalities*. Modalities are in fact functions. That is, if M is a modality and W is a context, $M(W)$ is the modal context[1] that corresponds to the point of view M contained in W. For instance, if $bel(A)$ is the modality of A's belief and W_1 and W_2 are modal contexts mentioned above, then $bel(A)(W_1) = W_2$ holds. Two kinds of modalities are important. One is names of states of the world or situations. They are treated as additional arguments of predicates in planning area [29], but can be viewed as a kind of modalities. Another kind of modalities is obtained from propositional attitudes or propositional functions, such as beliefs, intentions and tenses. From a propositional function F and a modal context W, we can get another context W' by collecting propositions P such that proposition $F(P)$ holds in W. In order to get a partial state of the world in this way, we require attitudes and propositional functions to be consistent and closed under logical consequence.[2] So for example, neither desires that can be inconsistent nor intentions in a certain sense [5] that are not closed under logical consequence can be used to get modalities.

Modalities can be *composed*. If M_1 and M_2 are modalities, we can compose them to get another modality M that satisfies the condition $M(W) = M_1(M_2(W))$ for all contexts W. For instance, we can compose two modalities, A's beliefs and B's beliefs to get another modality, namely, A's beliefs about B's beliefs. Modalities have various properties according to their types, and frameworks for representation of agent's mental states must have a strong descriptive power for modalities and composition of them.

3.4 Previous Approaches

The most widely used approach to modalities is that based on *modal logics*, where modalities are treated as modal operators, that is, propositional operators. Knowledges, beliefs [16] and tenses are formalized as kinds of modal operators, and dynamic logics [17] that deal with actions permit composition of modalities. In modal logics, however, modalities are not first class citizens of the system, and it is unable to unify two modalities, nor to quantify over the set of modalities. It is a serious defect in planning and dialogue processing. Moreover, modal logics do not support facilities for local reasoning, since they have no means to represent modal contexts themselves. Since all information is represented in one layer, the base context in our

[1] Modalities correspond to relative pathnames in hierarchical file systems, whereas modal contexts correspond to absolute pathnames.

[2] In terminology of modal logic [9], we only deal with *normal* modalities.

words, reasoning about more composite modal contexts, more complex and longer reasoning process are required. We think it is unnatural for a model of human reasoning ability on modalities.

Situations in situation theory [12] are partial descriptions of the world and are first class citizens of the theory. They can be used for various purposes in dialogue understanding and commonsense reasoning. Regarding the supports relation $\models$ between situations and infons as a relation dependent on situations it is in, we can identify situations with modalities. For example, we can express the fact "Pat knows that P" by $pk \models P$ where pk is a situation which expresses Pat's knowledge as in [31]. Although situation is a very powerful notion in situation theory, there exist no notion of joining two situations in the way parallel to composing two modalities. Without such a notion, we can use only the basic modalities, and hence are forced to use very complex expressions in order to describe their properties.

Chapter 4

Mental World Structures

We represent the agent's mental states in the form called *Mental World Structure(MWS)* which consists of multiple mental worlds. Each mental world is viewed as a set of mental propositions, basic units of information possessed by the agent, and corresponds to one modal context. With these representations of modal contexts, local reasoning is realized by context (world) relative definitions of the inference procedures given in Chapter 5. We represent modalities by *path expressions*, which are first class citizens of the framework and can be freely composed each other.

4.1 Syntactic Objects

4.1.1 Overview

First, we explain syntactic objects of our framework, namely, objects that can occur in mental propositions. Formally, they are classified into six types: (i)*relation*, (ii)*polarity*, (iii)*proposition*, (iv)*path expression*, (v)*object*, and (vi)*function*. The first four types are subtypes of the fifth type, object type, and they are all first class citizens of our representation, i.e., can be arguments of mental propositions. Mental propositions are objects of proposition type and usually take the form

$$\ll rel, a_1, \ldots, a_n; pol \gg$$

where rel is a relation, $a_1, \ldots, a_n$ are objects, and pol is a polarity. This proposition expresses information that objects $a_1, \ldots, a_n$ stand at a relation rel if polarity pol is `1`, and that they do not if pol is `0`. We call $a_1, \ldots, a_n$ the arguments of the proposition and pol its polarity. We represent possibly composite modalities by path expressions.

Mental propositions in our framework are very similar to *infons* in situation theory [12]. They can take as their arguments path expressions and situations, respectively, both of which are used to represent modalities. Hence, our representation framework is considered to be a (simple) version of situation theory that integrates a notion of composing modalities.

4.1.2 Symbols

We use the following four kinds of symbols besides parentheses and other auxiliary symbols:

- *parameters*: a, b, c, ...
- *fresh parameters*: $a, $b, $c, ...
- *variables*: X, Y, Z, ...
- *functions*: f, g, h, ...

Each parameter, fresh parameter and variable have their types which are one of the above first five types, and each function is of function type.

Parameters are the most basic syntactic objects in our framework, and denote atomic mental objects or individual concepts of an agent. For example, apple_1 might be a parameter of object type which denotes agent's mental object that corresponds to an apple in front of him, red is a parameter of relation type, and 1 and 0 are parameters of polarity type. In the representation, parameters act like usual constants except for equality conditions. A constant is usually equal to itself and to nothing else. On the contrary, it is often the case that an agent who formerly thought two objects were different comes to know they are the same, as his information grows. For instance, an agent may happen to know that apple_1, the apple in front of him, is identical to the apple, say apple_2, he knows his mother bought yesterday. In this case, he adds to his representation the following equality proposition

$$\ll \texttt{equal, apple\_1, apple\_2; 1} \gg$$

and identifies these two parameters afterward.

Fresh parameters and variables are introduced in relation to quantification, which we explain below. We also add function forms to our representations for convenience. Each function has its *index* $(\alpha_1, \ldots, \alpha_n; \beta)$, where $\alpha_1, \ldots, \alpha_n$ and β are types. It means that this function takes n arguments of types $\alpha_1, \ldots, \alpha_n$ respectively and then forms an object of type β. We express parameters and functions by atomic symbols beginning with small letters, fresh parameters by symbols beginning with the letter $, and variables by symbols beginning with capital letters. We also use meta-variables that denote syntactic objects, which are expressed by italic letters.

4.1.3 Syntactic Objects

Now we define objects of six types: (i)relation, (ii)polarity, (iii)proposition, (iv)path expression, (v)object, and (vi)function inductively by the following rules:

1. parameters, fresh parameters and variables are objects of their types;

2. if *rel* is a relation, $a_1, \ldots, a_n$ are objects, and *pol* is a polarity, then $\ll rel, a_1, \ldots, a_n; pol \gg$is a proposition;

3. if $p_1, \ldots, p_n$ are path expressions, then $p_1 \ldots . p_n$ is a path expression (in particular, . is a path expression);

4. relations, polarities, propositions and path expressions are objects of object type;

5. if f is a function with index $(\alpha_1, \ldots, \alpha_n; \beta)$ and $a_1, \ldots, a_n$ are of types $\alpha_1, \ldots, \alpha_n$ respectively, then $f(a_1, \ldots, a_n)$ is an object of type β.

Path expressions are used to express modalities. A composite path expression $p_1 \ldots . p_n$ expresses the modality obtained by composing n modalities which are expressed by path expressions $p_1, \ldots, p_n$ respectively. According to rule 3, each path expression has a tree structure. It is redundant, however, since we want to treat path expressions as mere sequences of basic ones. So afterwards we identify path expressions

$$p_1 \ldots . p_{i-1} . (q_1 \ldots . q_m) . p_{i+1} \ldots . p_n$$

with

$$p_1 \ldots . p_{i-1} . q_1 \ldots . q_m . p_{i+1} \ldots . p_n,$$

and in particular,

$$p_1 \ldots . p_{i-1} . (.) . p_{i+1} \ldots . p_n$$

with

$$p_1 \ldots . p_{i-1} . p_{i+1} \ldots . p_n.$$

4.1.4 Modalities and Implications

Various modalities are expressed by parameters and function forms of path expression type. For instance, modalities explained in the example in Chapter 3 are expressed by `bel(`A`)`, `int(i)` and `past_1`, where `past_1` is a parameter of path expression type, `i` is a parameter of object type which denotes the agent himself, and `bel` and `int` are functions with index *(object; path expression)*.

We express the fact that a proposition P holds in a point of view expressed by a path expression $PATH$ by the following proposition.

$\ll$`hold,` $PATH$`,` P`; 1`$\gg$

For example, we express the fact that an agent A believes that P by

$\ll$`hold, bel(`A`),` P`; 1`$\gg$.

Variables in propositions are considered to be universally quantified, as in logic programs. Fresh parameters in propositions are considered as Skolem constants or functions, and act like existentially quantified variables. So, the proposition

$$P(X_1, \ldots, X_n, \$a_1, \ldots, \$a_m)$$

in our framework corresponds to the formula

$$\forall X_1 \ldots \forall X_n \exists Y_1 \ldots \exists Y_m \, P(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$

in the ordinary form. Quantification has special importance in *implications*, that is, propositions of the form

$$\ll \texttt{imply},\ ANT,\ CON;\ \texttt{1} \gg$$

where relation `imply` takes as its arguments two lists[1] of propositions, the *antecedent list* and the *consequent list*, and expresses the fact that if all the propositions in the antecedent list hold then all the propositions in the consequent list hold. For instance, the implication

```
≪imply,
        [≪human, X≫],
        [≪age, X, $a≫]≫
```

[2]

expresses the fact which can be expressed by the ordinary first order formula

$$\forall X \exists Y \; human(X) \supset age(X, Y)$$

that is, the fact that every human has his own age.

4.2 Definition of the structure

A *mental world* is a path expression containing no fresh parameter and no variable. Mental worlds are used to represent modal contexts, and composite mental worlds represent composite points of view in the same way as composite path expressions represent composite relative points of view. In particular, mental world "." represents agent's beliefs of the most simple type, namely, the base context. We call this world the *base world*. We say that a world W is an *ancestor* of a world W' and W' is a *descendant* of W when W is a prefix of W', and in this case, we call path expression $W' - W$ *the path from world* W *to* W'. It expresses the point of view contained in W which world W' represents.

A *node* is a pair (W, P) where W is a mental world and P is a proposition. It is used to represent the fact that P holds in W. For example, two nodes

```
( . , ≪hold, bel(A), P; 1≫)
(bel(A) , P)
```

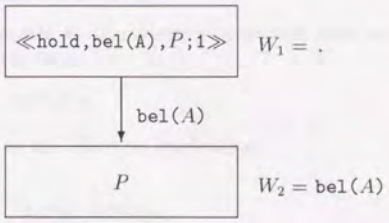


Figure 4.1: Equivalent Nodes

represent the same fact, the agent's believing that A believes proposition P (see Figure 4.1). In general, many different nodes can be used to represent one fact, so we define here an equivalence relation of nodes. At first, we define the *normal form* of a node by the following rewriting rules.

- $(W, \ll \texttt{hold},\ PATH.PATH',\ P;\ pol \gg) \rightarrow (W', \ll \texttt{hold}\ PATH',\ P;\ pol \gg)$ where $PATH$ contains no fresh parameter and no variable, and is the path from world W to W'
- $(W, \ll \texttt{hold},\ .\ \ ,\ P;\ 1 \gg) \rightarrow (W, P)$

Two nodes are said to be *equivalent* if and only if their normal forms are identical. Equivalent nodes represent just the same facts and are treated equally in inference. Now we define a *Mental World Structure* M to be a set of nodes $node(M)$ together with a binary relation C on the set $node(M)$. We say proposition P *holds in* a world W if $node(M)$ contains a node that is equivalent to (W, P). The binary relation C expresses *dependence* between propositions contained in the structure. If two nodes N, N' stand at this relation, we say that N' is *caused by* N, or that there is a *causality link* from N to N'. This dependency information is updated constantly through inference, and used to maintain dependency between nodes by truth maintenance inference procedure.

4.3 Examples

This section gives several examples that demonstrate strong descriptive power of our framework for modalities.

[1]Lists are considered to be constructed by **cons**, a function with index $(object, object; object)$.
[2]We often omit the polarity 1 of propositions.

4.3.1 Representation of Beliefs

It is often the case that a mental world *inherits* propositions from another world. We can express this fact by the form

$$\ll\texttt{inherit, } PATH\texttt{, } PATH'\gg$$

whose meaning is explained by the following implication.

```
≪imply,
        [≪inherit, PATH, PATH2≫,
         ≪hold, PATH, P; 1≫],
        [≪hold, PATH2, P; 1≫]≫                    (1)
```

That is, if path expression $PATH'$ inherits from $PATH$, a proposition solvable in the world referred to by $PATH$ is also solvable in the world referred to by $PATH'$. With this inheritance relation, we can express the fact that if an agent believes some proposition then he believes that he believes that proposition as follows:

```
≪inherit, bel(A), bel(A).bel(A)≫                  (2)
```

As for the agent's own beliefs, the agent thinks all his beliefs to be true, so the following proposition holds.

```
≪inherit, bel(i), . ≫                             (3)
```

Here, parameter `i` denotes the agent himself.

4.3.2 Commonsenses

Although (3) holds only in the base world, proposition (1), definition of a relation, and proposition (2), a general property of beliefs, are considered to hold in all mental worlds in the structure. We shall call such propositions *commonsenses*. In our framework, we can express that a proposition $PROP$ is a commonsense by adding the following proposition to the base world.

$$\ll\texttt{hold, X, } PROP\texttt{; 1}\gg$$

Hence for example, we express the property of beliefs mentioned above by adding the following proposition to the base world of the structure.

```
≪hold, X, ≪inherit, bel(A), bel(A).bel(A)≫; 1≫
```

Usual axioms and inference rules of the theories are all commonsenses in this sense. In our framework, we can express them in the same level as other facts about the current state of the world by quantifying over the set of path expressions.

4.3.3 Representation of Mutual Beliefs

As another example of quantification over modalities, we shall take mutual beliefs between two agents. Agents A and B have *mutual belief* that $PROP$ if and only if the condition

G_1 believes that G_2 believes that $\ldots G_n$ believes that $PROP$

holds for all positive integers n and all combinations of $G_1 \ldots G_n$ each of which is either A or B . We adopt here an indirect approach to mutual beliefs though they can be expressed directly as a path expression in the same way as private beliefs. We introduce a property `mbel(`A`,`B`)` of path expressions by the following commonsenses.

```
≪mbel(A,B), bel(A)≫
≪mbel(A,B), bel(B)≫
≪imply,
        [≪mbel(A,B), PATH≫],
        [≪mbel(A,B), PATH.bel(A)≫]≫
≪imply,
        [≪mbel(A,B), PATH≫],
        [≪mbel(A,B), PATH.bel(B)≫]≫
```

and express the fact that $PROP$ is mutually believed between A and B by

```
≪imply,
        [≪mbel(A,B), PATH≫],
        [≪hold, PATH, PROP≫]≫.
```

Chapter 5

The Inference Procedures

The problem of designing an inference mechanism can be divided into the following two subproblems.

- From the cognitive viewpoint, what kind of inference procedure is permitted to apply?
- Provided that we have chosen the set of inference procedures, how can we appropriately control or schedule them?

As for the first problem, we adopt three procedures as the basic inference procedures, that is, *deduction*, *abduction(hypothesis generation)*, and *truth maintenance*. They are applied to a MWS and transform it, namely, add or delete nodes and causality links. The agent's all inference processes are modelled by successive applications of these three procedures in an appropriate order.

We assume that applications of the procedures are scheduled by a certain inference control unit. The problem of inference control is in general very hard, and formulation of the unit is beyond the scope of this thesis, though we will give some further remarks about inference control in Chapter 8.

Our three basic inference procedures are defined relative to mental worlds, hence are applied equally to each mental world in MWS. Before presenting their precise description, we need to introduce several notions and conventions. We denote by $P[f]$ the proposition obtained from a proposition P by a substitution f of variables, and call it an instance of P. If S is a set of propositions, then we denote by $S[f]$ the set of propositions of the form $P[f]$ where $P \in S$.

We define a proposition P being *solvable in* a world W *on the basis of* a set of nodes BA inductively as follows.

1. if P holds in W, P is solvable in W on the basis of $\{(W, P)\}$.
2. if an implication $\ll$`imply`, ANT, $CON\gg$ holds in W and every proposition Q in $ANT[f]$ is solvable in W on the basis of a set of nodes BA_Q, then every proposition in $CON[f]$ is solvable in W on the basis of $\{(W, \ll$`imply`, ANT, $CON\gg)\} \cup (\cup_{Q \in ANT[f]} BA_Q)$.

3. if propositions P and $\ll$`equal`, OBJ, $OBJ' \gg$ is solvable in W on the basis of BA and BA' respectively, then propositions obtained from P by substituting some occurrences of OBJ with OBJ' are solvable in W on the basis of $BA \cup BA'$.

4. if P is solvable in W on the basis of BA, instances of P are solvable in W on the basis of BA.

The inference mechanism can easily check whether a proposition is solvable in a world or not in a backward-chaining manner.

The deduction and the abduction procedures add propositions to mental worlds in the structure, and add causality links to the structure. The truth maintenance procedure deletes them. Here and below, we mean by addition of a proposition P to a world W in a MWS M, addition of the *normal form* of node (W, P) to $node(M)$,[1] and mean by addition or deletion of causality links, extension or restriction of the dependency relation C.

5.1 Deduction

The *deduction* inference procedure uses implications forwardly to obtain new information. That is, when an implication holds in a mental world and all its antecedent propositions are solvable there, we can apply the deduction procedure to add each consequent proposition of the implication to that world. Note that most use of this procedure is dispensable, since we can get the same information by backward-chaining in other inference procedures. So deduction functions like *partial computation*, and deciding how much of it should be done is mainly a problem about computational efficiency.

When an implication

$\ll$`imply`, ANT, CON; `1`$\gg$

holds in a world W and for some substitution f every proposition Q in $ANT[f]$ is solvable in W on the basis of a set of nodes BA_Q, we can apply the deduction procedure. The procedure substitutes all occurrences of *fresh* parameters contained in propositions in $CON[f]$ with newly created parameters, and then adds resulting propositions to W. Moreover, it adds to the structure causality links from each node in $\{(W, \ll$`imply`, ANT, $CON \gg)\} \cup \cup_{Q \in ANT[f]} BA_Q$ to each added node.

Let us take an example. Consider an agent heard another agent `sp` uttering a sentence "John runs." As a result of recognition, the following proposition is added to the base world of his MWS.

$\ll$`done, utter(sp, [john, runs])`$\gg$ (1)

[1] We use the normal form to promote local reasoning.

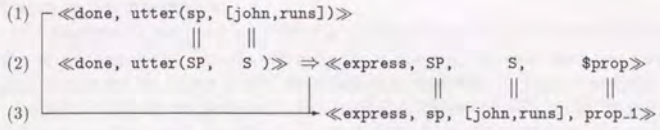


Figure 5.1: An Application of the Deduction Procedure

Here, relation **done** takes a name of an action as its argument, and means that that action has just been done. It is almost always true that if some agent utters a sentence then he expresses by it some proposition, so the base world also includes the following proposition.

```
≪imply,
        [≪done, utter(SP, S)≫],
        [≪express, SP, S, $prop≫]≫                    (2)
```

If he applies the deduction procedure in the base world, proposition

```
≪express, sp, [john, runs], prop_1≫                   (3)
```

is added to the base world and two causality links

$(\ .\ ,(1)) \rightarrow (\ .\ ,(3))$
$(\ .\ ,(2)) \rightarrow (\ .\ ,(3))$

are added to the structure (see Figure 5.1: the thick arrows denote implications and the thin arrows denote causality links). Here, **prop_1** is a newly created parameter and corresponds to his image of the meaning of the sentence. Through this inference, the agent understands that agent **sp** means some proposition by uttering the sentence "John runs," and will understand the meaning of it when he further succeeds to unify parameter **prop_1** with some definite proposition.

5.2 Abduction

An agent does not always draw sound inference. Under circumstances where necessary information is not given enough, he must generate hypotheses and then commit himself to them. In particular, when an agent has information without explanations such as an observed fact, he tries to generate a hypothesis which explains that information. Inference of this type is generally called *abduction* or *abductive reasoning*. We formulate abduction as one of the basic inference procedures of our framework, which adds as an explanation a set of propositions which proves a given proposition. To find such a set, it uses implications and unification between parameters.

Before giving a precise description of the abduction procedure, we briefly explain its significance in our framework.

In this framework, we must be able to understand agent's inference process purely logically or declaratively. On the other hand, it is often the case that the results of inference drawn by an agent is not determined completely in logical or declarative level, namely, without *preferences*. For instance, there usually exist many plans that achieve a given goal, so when an agent intends some proposition, we cannot predict which plan he chooses purely logically. Another example is disambiguation of linguistic expressions. In many cases, a given expression has several interpretations that satisfy all syntactic, semantic and pragmatic constraints. So we must be able to deal with this sort of nondeterminism in our framework, and for this purpose, we can use the abduction procedure which nondeterministically chooses an explanation of a given proposition. We present examples of application of abduction to reasoning about linguistic expressions and planning below and in the next chapter, respectively.

When we apply the abduction procedure to a proposition P that holds in a world W, in order to find an explanation of it, the procedure first substitutes every parameter PAR in P with a new variable X_{PAR}. We denote the resulting proposition by Q. Then it performs one of the following procedures.

1. if for some substitution f, $Q[f]$ holds in W, the procedure unifies P and $Q[f]$. That is, it adds equality propositions of the following form to W

   $$\ll \texttt{equal},\ PAR,\ X_{PAR}[f] \gg$$

   and then adds causality links from (W, P) and $(W, Q[f])$ to each added node.

2. if there exists an implication $\ll \texttt{imply},\ ANT,\ CON \gg$ holding in W such that $Q[f]$ is an element of $CON[f]$ for some substitution f, the procedure unifies P and $Q[f]$ in the same manner as above, substitutes all variables that occurs in $ANT[f]$ but not in $Q[f]$ with newly created parameters, and adds resulting instances of $ANT[f]$ to W. Moreover, it adds causality links from (W, P), $(W, \ll \texttt{imply},\ ANT,\ CON \gg)$ and each equality node added through unification to each added instance of $ANT[f]$.

Although we can apply the abduction procedure to arbitrary proposition P in the structure, two cases are particularly important. One case is: when P has no explanation yet, that is, is not proved from other propositions in the structure. Such a proposition may be introduced through recognition of the outer world or other application of the abduction procedure. Another important case is: when P contains a parameter created just now. In this case, application of the procedure often results in unifying this parameter with other object.

As an example for use of abduction, we shall continue the above linguistic example. Proposition (3) contains a newly created parameter `prop_1`, so the agent tries to find an explanation of this proposition. We assume

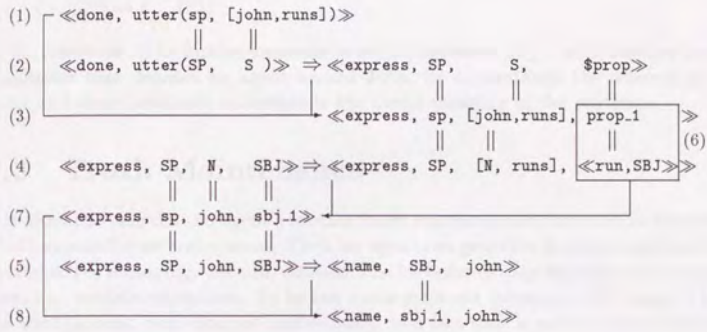


Figure 5.2: Interplay of Deduction and Abduction

```
≪imply,
        [≪express, SP, N, SBJ≫],
        [≪express, SP, [N, runs], ≪run, SBJ≫≫]≫        (4)
```

and

```
≪imply,
        [≪express, SP, john, SBJ≫],
        [≪name, SBJ, john≫]≫                            (5)
```

which roughly describe the meaning of the verb `runs` and the noun `john`, respectively. Applying the abduction procedure to (3) and (4), he adds

```
≪equal, prop_1, ≪run, sbj_1≫≫                           (6)
≪express, sp, john, sbj_1≫                              (7)
```

to the base world, and causality links

$$(\,.\,,(4)) \rightarrow (\,.\,,(7))$$
$$(\,.\,,(6)) \rightarrow (\,.\,,(7))$$

to the structure (see Figure 5.2). Newly created parameter `sbj_1` corresponds to his image of the referent of "John." Again, the deduction procedure is used to add

```
≪name, sbj_1, john≫                                     (8)
```

to the base world, and causality links,

$(\ .\ ,(5)) \rightarrow (\ .\ ,(8))$
$(\ .\ ,(7)) \rightarrow (\ .\ ,(8))$

to the structure. If he further succeeds to unify parameter `sbj_1` with another known parameter that denotes an agent named John, he understands the referent of the noun and simultaneously understands the literal meaning of the sentence.

5.3 Truth Maintenance

It is often the case that an agent's mental states representation becomes inconsistent. This happens for several reasons. First, an agent can generate incorrect explanations by abductive reasoning. Second, implications he believes may be in fact *prototypical* ones, i.e., contain exceptions. So he can make incorrect inference with them. Third, the world around him changes continuously, so a fact that is correct at one moment may become false in the next moment.

When the representation becomes inconsistent, we use the *truth maintenance* inference procedure, which gets rid of inconsistency taking dependency of propositions into consideration. That is, when a proposition P and its dual $\bar{P}$[2] are both solvable in a world W in the structure on the basis of sets of nodes BA and BA' respectively, this procedure chooses one base (W', Q) from $BA \cup BA'$, and deletes Q from the world W'. Deleting Q, the procedure checks whether there exist nodes that are caused by (W', Q), and if there exists such a node N, it deletes N next in the same manner. [3]

[2]In other words, P and $\bar{P}$ are the same propositions except for polarities being opposite.

[3]This is the simplest formulation of the truth maintenance task. More detailed formulation of it is left for the future work.

Chapter 6

Example Dialogues

Consider a clerk at the information desk in a station who is asked by a customer,

"I want to go to the museum."

If the clerk thinks that his customer should take a bus to go there and does not think that the customer knows where it starts, he should supply the customer that information by saying

"The bus starts from gate 3." (Response 1)

In another situation, where the clerk knows that the museum is closed that day and the customer cannot enter it even if he goes in front of it by bus, the following response is perhaps more helpful to the customer.

"Sorry, it is closed today." (Response 2)

Below we explain in our framework the clerk's inference processes for generating these two responses.

Cooperative dialogues of this sort have been treated by Allen [1] and many followers of plan-based approach to dialogue processing. However, they use special data structures for plans and special rules of inference about them, and hence it is hard to make those dialogue processing models more flexible by incorporating facilities for various other types of inference.

On the other hand, we explain the above dialogues in a general manner in our unified framework for problem solving. That is, both agents' beliefs and states of the world in the future are represented as kinds of modalities, and both plan construction and plan recognition processes are modelled by applications of the abduction procedure.

We do not here deal with interpretation and generation of linguistic expressions, and the clerk's inference processes described below begin with his recognition of the customer's performance of an abstract informing illocutionary act [40] with some propositional content and end with his own performance of another illocutionary

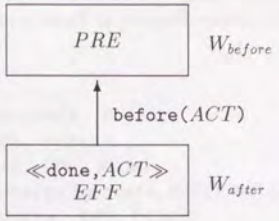


Figure 6.1: Representation of an Action

act.[1] In the end of this chapter, we briefly discuss how we can incorporate inference about linguistic expressions into our examples.

6.1 Commonsenses

We list here the clerk's background knowledge used later. They are all considered to be commonsenses, that is, hold in all mental worlds in his MWS.

Actions are regarded simply as transformations of states of the world in this chapter. If W_{after} is a mental world which represents the state of the world just after a performance of an action ACT, it includes proposition

≪done, ACT≫

and there we can refer by path expression before(ACT) to the mental world which represents the state just before that performance of ACT (we call this world W_{before}), as illustrated in Figure 6.1. ACT's precondition PRE holds in W_{before}, and its effect EFF holds in W_{after}. We express this fact by propositions of the following form.

≪imply,
[≪done, ACT≫],
[≪hold, before(ACT), PRE≫,
EFF]≫

We use three kinds of actions in our examples.

≪imply,
[≪done, inform(SP, HR, PROP)≫],
[≪hold, before(inform(SP,HR,PROP)).bel(SP), PROP≫,
≪hold, bel(HR).bel(SP), PROP≫]≫ (C1)

[1] We assume in this framework that an agent performs an action when he intends to perform it and all its preconditions are solvable in his MWS.

That is, an informing illocutionary act is performed only when the speaker believes its propositional content, and as a result of its performance, the hearer knows that the speaker believes the content.

```
≪imply,
        [≪done, go_by_bus(A, BUS)≫,
         ≪source, BUS, P1≫,
         ≪destination, BUS, P2≫],
        [≪hold, before(go_by_bus(A,BUS)).bel(A),
                ≪source, BUS, P1≫≫,
         ≪at, A, P2≫]≫                                    (C2)
```

An action of going somewhere by a bus is performed only when the actor knows where the bus starts, and after its performance, he is at the destination of the bus.

```
≪imply,
        [≪done, enter(A, PL)≫],
        [≪hold, before(enter(A,PL)), ≪at, A, PL≫≫,
         ≪hold, before(enter(A,PL)), ≪open, PL≫≫,
         ≪in, A, PL≫]≫                                    (C3)
```

An agent enters some place PL only when he is in front of PL and PL is open, and then he is in PL. We also use the following contraposition of C3.

```
≪imply,
        [≪open, PL; 0≫],
        [≪done, enter(A, PL); 0≫]≫                        (C4)
```

Intentions are treated as a kind of beliefs about the future. If an agent A intends to make proposition P true, he believes that P holds in some state of the world in the future, that is, two propositions

```
≪future, T≫
≪hold, T, P≫
```

hold for some path expression T in the mental world for A's beliefs. Here, T corresponds to a state of the world in a specific instant of time and may be either definite or indefinite. `future` is a property of path expressions which means that its argument represents a state of the world in a future instant of time. We also use property `today`, which means it argument corresponds to an instant of that day. We assume actions complete at once, hence the following implications all hold.

```
≪imply,
        [≪future, T≫],
        [≪future, T.before(ACT)≫]≫                        (C5)
≪imply,
```

```
        [≪future, T.before(ACT)≫],
        [≪future, T≫]≫                                    (C6)
≪imply,
        [≪today, T≫],
        [≪today, T.before(ACT)≫]≫                         (C7)
≪imply,
        [≪today, T.before(ACT)≫],
        [≪today, T≫]≫                                     (C8)
```

The following inheritance rule plays an important role in our explanation of the helpful responses.

```
≪imply, [≪hold, bel(A), P≫], [P]≫                         (C9)
```

If an agent believes that another agent believes proposition P, then he comes to believe P. This rule should not be applied when the agent firmly believes the dual of P already.

When an agent thinks that another agent has beliefs that conflict with his own, he may point out the difference. In particular, differences about actions they intend to do are so serious that the agent must point it out to behave cooperatively, hence we get our last commonsense.

```
≪imply,
        [≪hold, bel(A), ≪future, T≫≫,
         ≪hold, bel(A).T, ≪done, ACT≫≫,
         ≪hold, T, ≪done, ACT; 0≫≫],
        [≪future, $t≫,
         ≪hold, $t.bel(A).T, ≪done, ACT; 0≫≫]≫            (C10)
```

That is, if an agent thinks that another agent A intends to perform an action ACT and that it is impossible, he intends to let A know that fact.

Our treatment of actions, intentions and time in this chapter is too naive, though it suffices for the current purposes. Utilization of more sophisticated theories of rational action [10], speech acts [11] and temporal logic [2] in our framework is left for the future work.

6.2 Supplying Relevant Information

The clerk's inference process for generation of Response 1 is shown in Table 6.1.[2] The normal forms of inferred nodes are listed in order of application together with names of applied procedures (deduction or abduction) and relevant node numbers. The hierarchy of mental worlds used in this example is illustrated in Figure 6.2.

[2]We often omit arguments of action expressions in tables and figures.

	Mental World	Proposition	Procedure
1	.	≪done, inform(you, i, ≪hold,t_1,≪at,you,m≫≫)≫	Given
2	bel(you)	≪future, t_1≫	Given
3	.	≪future, t_1≫	De C9,2
4	before(inform).bel(you).t_1	≪at, you, m≫	De C1,1
5	bel(i).bel(you).t_1	≪at, you, m≫	
6	bel(you).t_1	≪at, you, m≫	De C9,5
7	t_1	≪at, you, m≫	De C9,6
8	t_1	≪done, go_by_bus(you, b_1)≫	Ab C2,7
9	t_1	≪source, b_1, p_1≫	
10	t_1	≪destination, b_1, m≫	
11	.	≪imply,[≪future,T≫],[≪hold, T, ≪source,b,gate3≫≫]≫	Given
12	.	≪imply,[≪future,T≫],[≪hold, T, ≪destination,b,m≫≫]≫	Given
13	.	≪equal, b_1, b≫	Ab 10,12
14	.	≪equal, p_1, gate3≫	Ab 9,11
15	t_1.before(go_by_bus).bel(you)	≪source, b, gate3≫	De C2,8,9,10
16	t_1.before(go_by_bus).bel(you) .bel(i)	≪source, b, gate3≫	Ab C9,15
17	t_1.before(go_by_bus)	≪done, inform(i, you, ≪source, b, gate3≫)≫	Ab C1,16
18	t_1.before(go_by_bus).before(inform).bel(i)	≪source, b, gate3≫	De C1,17

Table 6.1: Clerk's Inference Process for Response 1

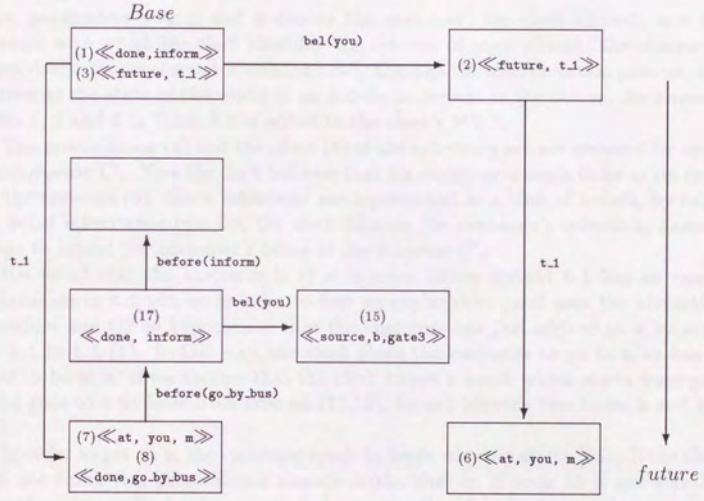


Figure 6.2: Mental Worlds for Response 1

The left half of the figure corresponds to the clerk's own beliefs and intentions, and the right half of it corresponds to his beliefs about the customer's beliefs and intentions.

When the clerk hears his customer's utterance of the sentence

"I want to go to the museum.",

he interprets it, and recognizes the following informing illocutionary act performed by the customer.

`inform(you, i, ≪hold, t_1, ≪at, you, m≫≫)`

Here, parameters `you`, `i` and `m` denote the customer, the clerk himself, and the museum with which the clerk identifies the referent of noun phrase "the museum", respectively. Parameter `t_1` is created newly through the interpretation process, and represents the state of the world in an indefinite instant in the future. As a result, nodes 1, 2 and 3 in Table 6.1 is added to the clerk's MWS.

The precondition (4) and the effect (5) of the informing act are deduced by using commonsense C1. Now the clerk believes that his customer intends to be at (in front of) the museum (6). Since intentions are represented as a kind of beliefs, by using the belief inheritance rule C9, the clerk inherits the customer's intention, namely, comes to intend the customer's being at the museum (7).

His belief that the customer is at `m` in some future instant `t_1` has no causal explanation in `t_1` yet, so he tries to find an explanation, and uses the abduction procedure and C2 to hypothesize that the customer has just arrived at `m` by some bus `b_1` in `t_1` (8). In this way, the clerk plans the customer to go to `m` by bus in order to be at `m`. If we assume that the clerk knows a bus `b` which starts from gate 3 and goes to `m` at least from now on (11,12), he can identify two buses `b` and `b_1` (13).

In order to get on `b`, the customer needs to know where it starts (15). If the clerk does not think this precondition already holds, that is, if node 15 is not solvable from the other nodes in the structure, he applies the abduction procedure again to provide an explanation of 15, and as a result, intends to perform the following action (17).

`inform(i, you, ≪source, b, gate3≫)`

The precondition of this illocutionary act (18) is solvable in the base world on the basis of nodes C5, C9, 3 and 11, so the clerk performs the act by uttering Response 1.

"The bus starts from gate 3."

To sum up, the clerk's inference process of constructing a plan for goal P is modelled here by successive applications of the abduction procedure to provide a causal explanation of the fact that P will hold in some future instant of time.

	Mental World	Proposition	Procedure
1	.	≪done, inform(you, i, ≪hold,t_1,≪at,you,m≫≫)≫	Given
2	bel(you)	≪future, t_1≫	Given
3	.	≪future, t_1≫	De C9,2
4	before(inform).bel(you).t_1	≪at, you, m≫	De C1,1
5	bel(i).bel(you).t_1	≪at, you, m≫	
6	bel(you).t_1	≪at, you, m≫	De C9,5
7	bel(you)	≪equal, t_1, t_2.before(a_1)≫	Ab C5,2
8	bel(you)	≪future, t_2≫	
9	bel(you).t_2	≪equal, a_1, enter(you, m)≫	Ab C3,6,7
10	bel(you).t_2	≪done, enter(you, m)≫	
11	.	≪today, t_1≫	Given
12	.	≪imply,[≪today, T≫], [≪hold, T, ≪open,m;0≫≫]≫	Given
13	t_2	≪open, m; 0≫	De 12,11,7, C8
14	t_2	≪done, enter(A, m); 0≫	De C4,13
15	.	≪future, t_3≫	De C10,8,10, 14
16	t_3.bel(you).t_2	≪done, enter(you, m); 0≫	
17	t_3.bel(you).t_2	≪open, m; 0≫	Ab C4,16
18	t_3.bel(you).bel(i).t_2	≪open, m; 0≫	Ab C9,17
19	t_3	≪done, inform(i, you, ≪hold, t_2,≪open,m;0≫≫)≫	Ab C1,18
20	t_3.before(inform).bel(i).t_2	≪open, m; 0≫	De C1,19

Table 6.2: Clerk's Inference Process for Response 2

6.3 Pointing out Customer's Plan Failure

Next, we consider the situation in which the museum m is closed that day. The clerk's inference process for generation of Response 2 is shown in Table 6.2 (see also Figure 6.3). Hearing the customer's utterance, the clerk infers in the same way as the preceding example, and believes that his customer intends to be at (in front of) the museum (6). Then he tries to find the reason why the customer has such an intention. Applying the abduction procedure, the clerk supposes that the customer intends to be at m just before some action a_1 is done (7,8), and that that action is his entering m (9,10). In short, the clerk now thinks his customer intends to be in front of m as the way to enter it (recognition of the customer's plan).

We assume that the clerk somehow knows the customer wants to go that day (11). Since the museum is closed in any instant which belongs to that day (12), he thinks that the customer's entering action can never be performed (14) which conflicts with his beliefs about the customer's beliefs. In order to resolve this conflict, the clerk

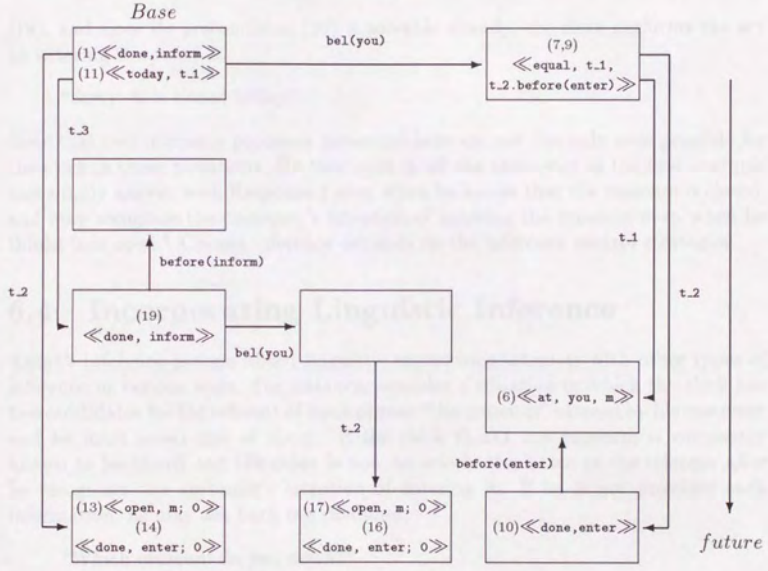


Figure 6.3: Mental Worlds for Response 2

forms an intention to let the customer know the impossibility of his action (15,16). Though the clerk can inform this information directly by uttering a sentence like

"You cannot enter it today.",

we assume here that he decides for some reason[3] to inform it indirectly by letting the customer know that the museum is closed that day (17). He intends to perform an informing illocutionary act

`inform(i, you, ≪hold, t_2, ≪open, m; 0≫≫)`

(19), and since its precondition (20) is solvable already, the clerk performs the act by uttering Response 2.

"Sorry, it is closed today."

Note that two inference processes presented here are not the only ones possible for the clerk in these situations. He may infer in all the same way as the first example and simply answer with Response 1 even when he knows that the museum is closed, and may recognize the customer's intention of entering the museum even when he thinks it is open.[4] Chosen inference depends on the inference control strategies.

6.4 Incorporating Linguistic Inference

Agent's inference process about linguistic expressions interacts with other types of inference in various ways. For instance, consider a situation in which the clerk has two candidates for the referent of noun phrase "the museum" uttered by his customer and he must select one of them. If the clerk thinks one museum is commonly known to be closed and the other is not, he selects the latter as the referent *after* he recognizes the customer's intention of entering it. If he is not provided such information, he may ask back the customer,

"Which museum do you mean?"

Our framework can be used to model such interactions of different types of inference process since it does not fix order of inference in advance and has sufficiently strong descriptive power to deal with both linguistic (syntactic, semantic and pragmatic) constraints and constraints about rational agents such as C9 and C10 above.

We have briefly illustrated in Chapter 5 linguistic inference process using propositions of the form

≪express, SP, EXP, SEM≫

[3]It depends on the inference control unit.

[4]In this case, however, the clerk still answers with Response 1 after all.

which means that a speaker SP expresses SEM by an utterance of a linguistic expression EXP. In our framework, various kinds of linguistic constraints are represented as constraints about this **express** relation, and we can then relate surface uttering acts with abstract illocutionary acts by rules such as

```
≪imply,
        [≪done, utter(SP, HR, S)≫,
         ≪express, SP, S, PROP≫,
         ≪declarative, S≫],
        [≪done, inform(SP, HR, PROP)≫]≫
```

though relation between them is in general very complicated and context-sensitive.

Now the preceding example is explained as follows. When the clerk tries to interpret the customer's utterance, he generates a new parameter `m_1` which corresponds to his image of the referent of noun phrase "the museum." If the clerk knows two museum one of which is open and the other is closed, he postpones unifying `m_1` with one of them, until he infers that the sentence uttered by the customer expresses proposition

```
≪hold, t_1, ≪at, you, m_1≫≫,
```

recognizes the performed illocutionary act

```
inform(you, i, ≪hold, t_1, ≪at, you, m_1≫≫),
```

recognizes the customer's intention of entering `m_1`, and then knows from C3 that the customer thinks `m_1` is open.

Chapter 7

Related Work

Our framework is in many aspects similar to a programming/knowledge representation language *PROSIT* [31, 39]. PROSIT is a language based on situation theory [12]. *Situations* are partial descriptions of the world and are first class objects in PROSIT. The supports relation $\models$ between situations and infons is treated to be dependent on situations it is in, hence one can suppose situations to be hierarchically structured, like mental worlds in our framework. PROSIT also support facilities for local reasoning, since a user can make queries in arbitrary situations in the hierarchy.

Pieces of information are represented by *infons*

$$(relation\ object_1\ \ldots\ object_n)$$

which are akin to our mental propositions. Parameters are also used in PROSIT as the basic syntactic objects, and they have the similar equality conditions as ours. That is, a parameter is only equal to itself through backward-chaining inference, but can be explicitly asserted to be equal to another object.

There are two inference procedures in PROSIT, namely, the *backward-chaining* and the *forward-chaining* inference. The backward-chaining procedure resembles that of Prolog using *backward-chaining constraints*

$$(\texttt{<=}\ head\ goal_1\ \ldots\ goal_n)$$

and is applied when a user inputs a query. The forward-chaining procedure is similar to our deduction procedure using another type of implications, *forward-chaining constraints*

$$(\texttt{=>}\ head\ result_1\ \ldots\ result_n)$$

and is applied whenever an infon matching *head* is asserted. Moreover a user can explicitly call one inference procedure during execution of the other procedure by built-in predicates.

The main difference between PROSIT and our framework is the way to *name* modalities (situations). In our framework, modalities are referred by any objects of

path expression type. In particular, we can use composite path expressions or functional forms such as `bel(`A`)` for names of modalities. In PROSIT on the other hand, situations are referred only by simple parameters. Without facilities of composing modalities (situations), PROSIT forces its user to use very complex expressions in order to describe relations among situations which are far from each other in the hierarchy. Perhaps the most serious problem is that PROSIT cannot deal properly with quantification over the set of situations in a simple manner. For instance, we have used in the last chapter the commonsense

```
≪hold, X, ≪imply, [≪hold, bel(A), P≫], [P]≫≫        (C9)
```

in several ways:

1. unifying `X` with ., deduce `≪future, t_1≫`
   from `≪hold, bel(you), ≪future, t_1≫≫`
   (Step 3),

2. unifying `X` with `t_1`, deduce `≪hold, t_1, ≪at, you, m≫≫`
   from `≪hold, bel(you).t_1, ≪at, you, m≫≫`
   (Step 7), and

3. unifying `X` with `bel(you).t_1`,
   deduce `≪hold, bel(you).t_1, ≪at, you, m≫≫`
   from `≪hold, bel(i).bel(you).t_1, ≪at, you, m≫≫`
   (Step 6).

Among these three deductions, only the second one can be done in PROSIT, since a situational variable unifies only with a situational parameter. Hence, commonsenses cannot be represented in PROSIT in the same way as described in this thesis.

As another example, consider the following discourse:

"Pat entered the museum. He went there by bus."

In our framework, the meaning of the two sentences are expressed by

```
≪hold, t_1, ≪done, enter(pat, m)≫≫
≪hold, t_1.before(enter(pat, m)), ≪done, go_by_bus(pat, b_1)≫≫
```

In general, we can uniformly express the meaning of past tense sentences by propositions of the form

`≪hold, `T`, `$PROP$`≫≫`

where T is a past instance of time (reference time) and $PROP$ is a tenseless proposition. In PROSIT on the other hand, the above discourse is expressed by

```
(!= t_1 (done (enter pat m)))
(!= t_1 (!= (before (enter pat m)) (done (go_by_bus pat b_1))))
```

Two infons have quite distinct forms, and there exists neither uniform schema for expressing past tense sentences nor concept of reference time.

Another difference between PROSIT and our framework is the inference mechanism. As a programming language, inference in PROSIT is mainly based on query answering. On the other hand, our framework is a general cognitive framework for an agent's problem solving, and has a set of the basic inference procedures that suffices for modelling the agent's inference processes in most of cooperative natural language dialogues. Our inference mechanism is powerful and general one, but as a result, we are imposed a hard problem, namely, the problem of inference control.

Chapter 8

Discussion: Controlling Inference

In Chapter 5, we have described the three available basic inference procedures, that is, deduction, abduction and truth maintenance, and assumed that applications of them are appropriately controlled and scheduled by some inference control unit. We have not designed such a unit completely, but here we discuss some idea about what it should be like.

The principle of local reasoning applies to inference control, also. That is, as the basic inference procedures are defined relative to mental worlds, the inference control unit applies equally to each mental world in the hierarchy. So, for example, a control unit which always gives high priority to inferences about the base world does not model an agent's reasoning ability correctly.

An inference control unit performs the following tasks:

1. determining which procedure is applied first,
2. selecting the best explanation of a given proposition in the abduction procedure, and
3. selecting the most appropriate proposition to delete in the truth maintenance procedure.

The first task is very important in our framework, since we are allowed to apply our general abduction procedure freely to arbitrary propositions. We need to restrict ourselves to inferences relevant to the agent's interests. There exist several heuristic strategies. First, applications of the abduction procedure to find an explanation of a proposition that has no explanation yet or contains parameters created just now should be given higher priority than other applications of abduction. Second, some of implications represented in the agent are mainly used in a forward-chaining manner. Deduction with such implications should be done whenever possible. Third, there exist domain-specific associative relations between propositions. For example, hearing an utterance, the hearer usually tries to recognize speaker's intention behind the utterance. Moreover, it is often the case that inferences indirectly related to other relevant inferences are also relevant to the agent. For instance, if an agent

believes ≪p, u≫ where u is a newly created parameter, then a deduction which concludes ≪p, a≫ for some object a may be relevant to him. Such an indirect relation between inferences can be formulated using a kind of spreading activation mechanism [18, 8].

The second and the third tasks of an inference control unit demand representation of preferences among candidates, i.e., possible explanations or propositions to delete. Preferences are determined by two factors, plausibility and utility. As an illustration of this, recall the example presented in Chapter 5. Hearing an utterance of noun phrase "John", the hearer infers that the referent of it (sbj_1) is an agent named John.

≪name, sbj_1, john≫

If he knows such an agent, say john_1,

≪name, john_1, john≫

he can explain parameter sbj_1 by unifying it with john_1.

≪equal, sbj_1, john_1≫

This explanation is preferred when it is plausible, i.e., he does not know any other Johns, or when it has high utility, for instance, when he wants to identify anyhow the referent of "John" in order to continue the dialogue. Which factor influences more depends on the problem area. In recognition problems such as utterance interpretation and plan recognition, plausibility is more influential than utility. On the other hand, in planning area both factors are equally important to select the best plan.

Preferences can be represented by numerical costs [20] or more symbolically [37]. Furthermore, we can formulate more complicated mechanisms to calculate preferences by using meta-level problem solving [34]. Incorporating appropriate representations of preferences into our framework is an important future subject.

Chapter 9

Conclusion

We have presented a new formal framework for problem solving of an intelligent agent who participates in a dialogue. We have given a precise definition of the representation structure and the basic inference procedures. Then we used this framework to explain agent's inference process in cooperative dialogues.

The main contributions of Part 1 are as follows:

1. We have presented a new formal framework for representing mental states of an agent called Mental World Structure, which has strong expressive power for modalities. Modalities can be composed, quantified and unified, and various types of knowledge that are difficult to express in previous representation systems have been shown to be expressed concisely in our framework.

2. We have smoothly incorporated into our framework the three basic inference procedures, that is, deduction, abduction and truth maintenance. They are defined relative to mental worlds, hence are applied equally to each mental world.

3. We have provided an explanation of an agent's inference processes working behind example cooperative dialogues, and demonstrated the strength of our representation system and inference procedures.

Part II

A Preferential Logic of Mental Attitudes

Chapter 10

Introduction

An agent who participates in a dialogue usually deals with problems that have more than one possible solution. He uses various kinds of preference to choose one or several most preferred solutions from possible ones, and acts according to his choice. Let us illustrate this briefly.

First, an agent understands his companion's utterance, which is generally ambiguous in some sense, and has several possible interpretations. For example, consider the companion's utterance

"I want to go to the museum."

It has the referential ambiguity for the noun phrase "the museum", since there exist more than one museum. The agent uses a preference about plausibility of interpretations that the nearest museum is most likely to be referred, and chooses the nearest museum as the referent of the noun phrase.

Next, to behave cooperatively, the agent recognizes his companion's intentions and plans from the utterance. Not only one plan can be ascribed to explain the given utterance. In our example, one possible plan consists of only one intention, the companion's intention to go in front of the museum. Another possible plan further includes his intention of entering it. In most cases, the agent chooses the latter plan and ascribes it to his companion, using a preference about plausibility that a person who wants to go in front of the museum usually wants to enter it.

Finally, the agent constructs a plan to make a response. There exist many and possibly infinite plans for him and sentences to utter. Preferences used here to choose a plan are not ones about plausibility as is used above but ones about desirability of responses. Using preferences for simple and helpful responses, he chooses and utters a response, for example,

"Sorry, it is closed today."

Note that this view of an agent's problem solving based on preferences gives a clear account of defeasibility of inferences, which is one of the most important features of human problem solving. As time goes by and a new piece of information is

obtained, solutions previously possible may become impossible or implausible. The set of the most preferred solutions changes, and thus an agent's previous conclusion may be abandoned.

In the traditional model of a dialogue participant, preferences are treated separately in each module for the dialogue task. In some models [1, 52], a part of preferences are represented explicitly by the evaluation rules. In the other models, preferences are generally not taken seriously and represented implicitly in the inference procedures. Semantic interpretation models [19, 38] use the specific procedures to choose plausible interpretations and referents. Plan recognition models [6] have procedures that find plausible plans of other agents, and linguistic generation models [4, 22] have procedures to generate helpful and clear responses.

However, such a module-specific or domain-specific treatment of preferences obstructs flexibility and clarity of dialogue participants' models. First, it is difficult to deal with interactions between preferences separately represented in different modules. It is often the case that a solution preferred in one module is defeated by another solution in another module. Let us take the problem of choosing the referent of the noun phrase "the museum" in the above example. In the semantic interpretation module the nearest museum is preferred. But, if the speaker's intention of entering "the museum" is inferred in the plan recognition module and the speaker is believed to know that the nearest museum is closed that day, another open museum may be preferred to the nearest closed museum in that module. To decide which museum is referred to, these two modules must interact each other.

Moreover, like logical constraints about the discourse domain, a single preference can be used in several modules. A linguistic preference, for example, that the referent of the expression should be obvious to the hearer, can be used in semantic interpretation and linguistic generation. A preference about actions, for example, that the speaker goes to a closed museum is generally not desirable, can be used in the plan recognition module and the plan construction module. Representing a single preference separately in different form in several modules makes a model unclear and a dialogue system hard to maintain. Therefore, we need a formal general framework for explicit representation of preferences. In fact, such a framework can also serve as a new basis for logics of mental attitudes, which is another approach to modeling a dialogue participant.

Mental attitudes are notions such as belief, knowledge and intention, which are attributed to agents. Logical analysis of these notions is a particularly important subject for dialogue processing, as well as other AI areas such as multi-agent systems, and it is intensively studied in recent years [10, 16, 27, 44]. Nevertheless, existing work only deals with more or less restricted phenomena, and the gap between logical theories and procedural models is still wide. One of the reasons of this is, in my opinion, that it ignores preferential aspects of mental attitudes. An agent forms his belief on the basis of his preferences about plausibility, and adopts intentions and plans based on his preferences about desirability. Moreover, the notion of preference itself can be considered to be a kind of mental attitudes. Analyzing mental attitudes

by preference will lead us to a natural and powerful theory of mental attitudes, and in particular, a proper treatment of the dynamics and interactions of attitudes.

In this thesis, we take this approach, and propose a preferential logic of mental attitudes. We deal with qualitative preferences, which are explicitly represented by partial orders on model structures. An agent's mental state is specified by knowledge and two preference orders, that is, the plausibility order and the desirability order. The language of our logic is a propositional language extended by adding attitudinal operators: belief, intention, choice, and preference between sentences. The satisfaction relation for these operators is defined in terms of the preference orders. Besides mental attitudes about the states of the world, mental attitudes about another agent's mental states can be dealt with. Furthermore, we introduce a construct of sentences, which is used to specify an agent's knowledge and the preference orders. Then we apply this logic to reasoning about plans. We give a formal account of plan construction process and examine several heuristics for plan recognition currently used.

Chapter 11

Need of Preference

11.1 Typology of Preference

We take the word "preferences" in a general and wide sense and imply an agent's evaluations for propositions or possible worlds. Here, each possible world is considered to be expanded in temporal order. By specifying a possible world, we specify a complete history of the world. We begin with discussions on the typology of preferences.

First, preferences are classified by their objects: those for propositions and those for possible worlds. Preferences for possible worlds are clearer in their meaning. An instance of them is a preference for a world w_1 to another world w_2. On the other hand, preferences for propositions are those such as a preference of a proposition ϕ to another proposition ψ. It is this type of preferences that is expressed by daily natural language utterances like

(1) "It is likely to rain tomorrow."
(2) "I prefer sushi to soba."

Sentence (1) expresses the speaker's preference for its raining the next day (to its being fine or snowing). Sentence (2) expresses the speaker's preferences for occurring his action of eating sushi to occurring his action of eating soba. We must be careful in dealing with preferences for propositions, since their real meanings are somewhat ambiguous. Regarding propositions as collections of possible worlds, we can reduce preferences for propositions to those for possible worlds. One possible reduction of a preference for a proposition ϕ to another proposition ψ is that possible worlds that satisfy ϕ are collectively or averagely preferred to those that satisfy ψ. Sentence (1) has this reading and can be interpreted as expressing that the probability of its raining tomorrow is high. Another possible reduction of the preference for ϕ to ψ is that each world that satisfies ϕ is preferred to each world that satisfies ψ. In this case, there is a further ambiguity whether the parts of the world other than ϕ and ψ must be fixed or may vary through the comparison. Sentence (1) also has this

1.	Object:	propositions or possible worlds
2.	Content:	plausibility or desirability
3.	Possessor:	own or other agents'
4.	Form:	numerical or symbolical

Table 11.1: Typology of Preference

reading and can be used to assert that the conditional probabilities of raining are high for all or certain many conditions. Sentence (2) has the latter reading only.

Second, we can distinguish preferences by what they are about. In this respect, an agent uses two types of preferences, one is about plausibility and the other is about desirability. Plausibility preferences are epistemological and particularly used in interpretative tasks. They serve as a supplementary role to an agent's knowledge which is always incomplete. An agent can judge whether a proposition (or a world) is plausible or not even when he does not know its actual truth. On the other hand, desirability preferences are related to an agent's behavior in general, and to generative tasks in particular. An agent evaluates the desirability of possible worlds, and plans and acts to make desirable worlds true.

The third classification concerns the possessors of preferences. In addition to reasoning with his own preferences, an agent can reason about other agents' preferences. Using knowledge about other agents' preferences, an agent recognizes their plans and predicts their actions. Most of daily preferences are shared by agents. Hearing the thunder rumbling, agents think it plausible to begin to rain, and agents think it undesirable to go out in the rain without an umbrella. However, agents may differ in their liking, and we assume that different agents generally have different preferences even when their states of knowledge are equal.

Fourth, preferences are represented in two distinct forms: quantitative approaches to preferences use their numerical representations and qualitative approaches use symbolical representations. The most widely known instance of the former approach is subjective Bayesian decision theory [26, 42]. Plausibility preferences are represented by a subjective probability function, and desirability preferences are represented by a utility function. Both functions assign to each possible world a real value, the probability and the utility of that world. Some AI systems use the similar numerical representations such as certainty factors [45] and assumability costs [20], though meanings of those values are comparatively vague. The key assumption of the qualitative approach is the totality of evaluation functions. Their values are linearly ordered and additive in the sense that they can be added and subtracted. On the other hand, the qualitative approach to preferences focuses only on their ordering, whose linearity is not necessarily assumed. Preferences are represented by an order relation on propositions or possible worlds. Order relations are usually assumed to be linear in decision theory, whereas partial orders are recently used to

represent plausibility of models in the study of nonmonotonic logics [37, 43].

11.2 Belief and Preference

Beliefs of an agent are consistent with each other and closed under logical consequence. That is, an agent never believes both a proposition ϕ and its negation $\neg\phi$ simultaneously, and an agent believes (at least implicitly) all consequences of his beliefs. Possible world model of belief [16] is used to formulate such static aspects of belief. An agent's state of belief is identified with the set B of all possible worlds that can be actual in the light of that state of belief. A proposition is said to be believed if it is true in all elements of B.

In addition to static properties, beliefs have dynamic properties. In particular, beliefs are nonmonotonic. An agent may abandon his beliefs when he obtains a new piece of information that contradicts with them. In terms of possible worlds, the set B may not be monotonically reduced along the passage of time. The dynamics of belief is separately studied as a theory of belief revision [14] , where postulates for revision are proposed.

To deal with the dynamics properly in possible world models, we need to introduce a notion of preference about plausibility, which ranks possible worlds. Then the set B is identified with the set of most preferred worlds that are consistent with knowledge. This idea is by no means new. Shoham and Cousins [44] presented the similar arguments, and Satoh [37] examined the relationship between possible world model with partial orders and the theory of belief revision.

11.3 Intention and Preference

Intention is an important notion when we consider communication between agents. An agent recognizes his companion's intentions to predict his actions and to behave cooperatively. Conversely, an agent indicates his intentions to his companion by actions. Natural language is a sophisticated tool for expressing intentions. Like beliefs, intentions have both static and dynamic properties. Intentions of an agent are consistent with each other, but they are not necessarily closed under logical consequence. That is, an agent does not need to intend all consequences of his intentions. Intentions are also nonmonotonic, but they have a certain persistency. Once an agent adopts an intention, he never drops it without special reasons.

Formal models of intention are proposed in recent years. For instance, Cohen and Levesque [10] analyzed intention based on a notion of persistency, and Konolige and Pollack [25] investigated a representationalist approach. However, none of the models is entirely satisfactory so far. Some of them establish counterintuitive assumptions, and the others fail to satisfy desirable properties such as ones noted above.

What is missing in previous work is consideration of the close relationship between intentions and preferences about desirability. This relationship is obvious: an agent generally adopts intentions in accordance with his preferences. A preference-based analysis of intention will help us to deal particularly with the dynamics of intentions and the relationship to other attitudes such as belief, choice and desire.

11.4 Reasoning about Plans and Preference

An agent constructs plans for the most desirable possible worlds and acts according to them. In the literature of plan construction and plan recognition, desirability is not directly dealt with. Rather a notion of goal is used as primitive, and plans for goals are studied.

There are a large number of plan construction models and planning systems [3, 7, 15]. Most of them only concern with providing plans that can be adopted by an agent for given goals rather than with determining what plans are really adopted. To determine what plans are adopted, we need preferences about desirability. An agent chooses the most desirable plans from many possible ones. Furthermore, if he gets new pieces of information, his preferences for plans may change, and then his plan may be revised. For example, consider an agent planning to go to a sushi bar to eat lunch. If he happens to know the bar is crowded, he may revise his plan and go to another restaurant, say, a noodle shop. Such phenomena must be explained in any satisfactory model of an agent.

Procedural models of plan recognition in dialogue understanding [1, 6] infer plausible plans of another agent from observed actions or intentions of that agent by using heuristic rules and inference procedures. Their theoretical foundations should be given by formal models of plan recognition, but existing formal models such as Kautz's [23] are not sufficient for this role. It is because they have no means to specify preferences and use only restricted type of preferences that come from structural properties of the domains like a preference for plans that consist of fewer actions. An instance of preferences they can not deal with is: it is plausible that an agent does not prefer eating in crowded restaurants. To model plan recognition properly, we need a general framework for representing desirability preferences of the planner, as well as the recognizer's plausibility preferences for mental states of the planner.

11.5 Our Approach

Our approach is based on two decisions. Our first decision is: we take a qualitative approach to preferences, treating preferences, both about plausibility and about desirability, simply as ordering, not numbers. There are two reasons for this decision.

First, we are skeptical about the totality of evaluation functions. It is doubtful that an agent assigns a definite value to every possible world (or proposition), since such thorough evaluation is hard to do and not necessary for him. In fact, we doubt

further the linearity of preference ordering. With two similar worlds, an agent may judge which one is preferred, or he may judge they are indifferent. But, an agent deals generally with worlds that have little in common. It is particularly the case when he reasons about another agent's mental states. With such a pair of worlds, he may not even think of comparing them. From a practical point of view, assigning values or linear orders to worlds in a consistent way is a troublesome task. On the other hand, it is convenient if we can deal with an agent's reasoning with partial information about preferences.

Second, we want to abstract the additive character of preferences to focus our interest on their ordering. It is ordering of preferences that directly influences an agent's attitudes and behavior. The most plausible worlds generally determine an agent's belief, and the most desirable worlds generally determine his plans and actions. Abstracting additivity of preferences, we lose composite notions such as probability of propositions calculated from probability of worlds and expected utility of actions [26], which are useful to select propositions and actions from those tying in the most preferred worlds. Nevertheless, as the first approximation of a full theory of an agent, we study an agent's qualitative reasoning about preferences.

Our second decision is: we use preference for model structures rather than use preference for propositions, which is a somewhat ambiguous notion, as a central semantic component of our logic. Here, a model structure consists of a possible world and a representation of another agent's mental state. This enables us to deal with preferences for the other agent's mental states, as well as preferences for possible worlds. Preferences are represented by two strict partial orders on model structures: the plausibility order represents plausibility preferences, and the desirability order represents desirability preferences. Note that these two orders are generally not related with each other. Then an agent's mental state is specified by three things: the plausibility order, the desirability order, and knowledge, which is represented by a set of model structures.

Preference for sentences is considered to be a kind of mental attitudes. We define it as well as other attitudes such as belief and intention in terms of the preference orders on model structures. Conversely, we give a method of specifying a preference order on model structures by a set of preferences between sentences. Then we apply this logic to reasoning about plans. We give a formal account of plan construction and plan revision processes, and we examine several heuristics for plan recognition currently used.

Chapter 12

Belief, Choice and Preference

12.1 Syntax and Semantics

In this section, we give a full syntax and semantics of our logic of mental attitudes. In this logic, we express mental attitudes of a dialogue participant called *Agent A*: attitudes about a planning domain and attitudes about mental attitudes of his companion, called *Agent B*, about that domain. Therefore, the language of our logic is three-layered. The bottom layer is a propositional language $\mathcal{L}$, by which the planning domain is described. The second layer consists of *B-sentences*, by which we express mental attitudes of Agent B. Finally, the top layer consists of *A-sentences*, by which we express mental attitudes of Agent A. We give these layers, together with their semantics, in turn.

12.1.1 A Propositional Language with Time Function

The language of our logic is based on a standard propositional language $\mathcal{L}$, which consists of a set of atoms $atom(\mathcal{L})$, $\wedge$, $\neg$, and other defined connectives. We denote atoms of $\mathcal{L}$ by $p, q, r, \ldots, a, b, c, \ldots$, and sentences of $\mathcal{L}$ by $\alpha, \beta, \gamma, \ldots$. We write $atom(\alpha)$ to mean the set of all atoms occurring in α. We express in $\mathcal{L}$ facts about a planning domain, such as some property's truth at some time and some action's occurrence at some time. To deal with a notion of time, We use a *time* function for $\mathcal{L}$ that assigns an element of a set $\mathcal{O}$ to each atom of $\mathcal{L}$, where $\mathcal{O}$ is linearly ordered by a temporal precedence relation $\leq$. We usually use the set of integers Z (with its standard numerical ordering) for $\mathcal{O}$. In those cases, the value of the time function is displayed in the name of an atom. For example, a sentence

$$eatA_1 \supset \neg hungryA_2$$

expresses the fact that if Agent A eats something at time 1 then he is not hungry at time 2, where $time(eatA_1) = 1$ and $time(hungryA_2) = 2$. An agent name is omitted from the name of an atom when it is clear from the context.

A *world* w is a function from $atom(\mathcal{L})$ to $\{true, false\}$. The satisfaction relation $w \models \alpha$ is defined in a usual way as follows:

Definition 12.1

1. For $p \in atom(\mathcal{L})$, $w \models p$ iff $w(p) = true$.
2. $w \models \alpha \wedge \beta$ iff $w \models \alpha$ and $w \models \beta$.
3. $w \models \neg\alpha$ iff $w \not\models \alpha$.

If $w \models \alpha$, we call w a model of α.
If $w \models \alpha$ for all worlds w, we write $\models \alpha$.

12.1.2 B-Sentences and B-Structures

B-sentences are used to express objects of Agent A's mental attitudes, that is, facts about the planning domain and Agent B's attitudes about it. To express mental attitudes, we introduce *attitudinal operators* for B: BEL_B, CHO_B, $P\text{-}PREF_B^T$, $D\text{-}PREF_B^T$, INT_B, SBG_B and $GINT_B$.

Definition 12.2

1. Atoms of $\mathcal{L}$ are *B-sentences.*
2. If α and β are sentences of $\mathcal{L}$ and $T \subset atom(\mathcal{L})$, then $BEL_B(\alpha)$, $CHO_B(\alpha)$, $P\text{-}PREF_B^T(\alpha, \beta)$, $D\text{-}PREF_B^T(\alpha, \beta)$, $INT_B(\alpha)$, $SBG_B(\alpha, \beta)$, and $GINT_B(\alpha)$ are *B-sentences.*
3. If ϕ and ψ are *B-sentences*, then $\phi \wedge \psi$ and $\neg\phi$ are *B-sentences.*

We denote B-sentences by $\phi, \psi, \chi, \ldots$. The intended meanings of attitudinal operators are as follows:

$BEL_B(\alpha)$	B believes α.
$CHO_B(\alpha)$	B chooses α.
$P\text{-}PREF_B^T(\alpha, \beta)$	B plausibly prefers α to β if all atoms in T are equal.
$D\text{-}PREF_B^T(\alpha, \beta)$	B desirably prefers α to β if all atoms in T are equal.
$INT_B(\alpha)$	B has an intention α.
$SBG_B(\alpha, \beta)$	B thinks that α is a subgoal of β.
$GINT_B(\alpha)$	B has a generalized intention α.

Model structures for B-sentences are called *B-structures.* Before presenting their definition, we need several notions for orders. A *strict partial order* is an irreflexive

and transitive relation. The set of *maximal* and *minimal* elements of a set U with respect to a strict partial order $<$ are defined as follows:

$$\begin{aligned} Max(U,<) &= \{x \in U \mid \text{there is no } y \in U \text{ such that } x < y\} \\ Min(U,<) &= \{x \in U \mid \text{there is no } y \in U \text{ such that } y < x\}. \end{aligned}$$

We say a strict partial order $<$ is *bounded* in a set U if for all nonempty $U' \subset U$, both $Max(U',<)$ and $Min(U',<)$ are nonempty.

Definition 12.3 $M = (W, \prec_{BP}, \prec_{BD}, w_0)$ is a *B-structure* iff

1. W is a nonempty set of worlds,
2. $\prec_{BP}$ and $\prec_{BD}$ are strict partial orders on worlds which are bounded in W, and
3. $w_0 \in W$.

The first three constituents of a B-structure specify Agent B's mental state, and the last constituent repersents an "actual" state of the domain. W specifies B's *knowledge*; it is the set of all worlds that can be actual in the light of his knowledge. The truth of B's knowledge is ensured by Condition 3 of the definition. Orders $\prec_{BP}$and $\prec_{BD}$are B's preference orders: the *plausibility order* and the *desirability order*, respectively. $w_1 \prec_{BP} w_2$ means that B plausibly prefers w_2 to w_1, or in other words, that B thinks w_2 is more plausible than w_1. If we assume B's subjective probability function P_B on worlds, $w_1 \prec_{BP} w_2$ implies $P_B(w_1) < P_B(w_2)$. The converse of this implication does not necessarily hold, since the plausibility order is partial and directly influences mental attitudes of an agent. Similarly, $w_1 \prec_{BD} w_2$ means that B desirably prefers w_2 to w_1, or in other words, that w_2 is more desirable for B than w_1. If U_B is a utility function for B, $w_1 \prec_{BD} w_2$ implies $U_B(w_1) < U_B(w_2)$. We impose the condition that the preference orders are bounded in order to simplify definitions and resulting properties when we formulate mental attitudes using maximally (minimally) preferred worlds. Note that when $atom(\mathcal{L})$ is finite, there are only finitely many worlds, and thus this condition is automatically satisfied.

Now we define the satisfaction relation for B-sentences. Propositional sentences are evaluated with respect to w_0, the actual state of the domain. Satisfaction of attitudinal sentences for B is defined in a completely parallel way with satisfaction of those for Agent A, so we give no further explanation of it here.

Definition 12.4 Let $M = (w_0, W, \prec_{BP}, \prec_{BD})$ be a B-structure.

1. For $p \in atom(\mathcal{L})$, $M \models p$ iff $w_0(p) = true$.
2. $M \models BEL_B(\alpha)$ iff $w \models \alpha$ for all $w \in Max(W, \prec_{BP})$.
3. $M \models CHO_B(\alpha)$ iff $w \models \alpha$ for all $w \in Max(Max(W, \prec_{BP}), \prec_{BD})$.

4. $M \models P\text{-}PREF_B^T(\alpha, \beta)$ iff $w_1 \prec_{BP} w_2$ for all pairs $w_1, w_2 \in W$ such that
   (a) $w_1 \models \neg\alpha \wedge \beta$,
   (b) $w_2 \models \alpha \wedge \neg\beta$, and
   (c) $w_1(p) = w_2(p)$ for all $p \in T$.
5. $M \models D\text{-}PREF_B^T(\alpha, \beta)$ iff $w_1 \prec_{BD} w_2$ for all pairs $w_1, w_2 \in Max(W, \prec_{BP})$ such that
   (a) $w_1 \models \neg\alpha \wedge \beta$,
   (b) $w_2 \models \alpha \wedge \neg\beta$, and
   (c) $w_1(p) = w_2(p)$ for all $p \in T$.
6. $M \models INT_B(\alpha)$ iff
   (a) $w \models \alpha$ for all $w \in Max(Max(W, \prec_{BP}), \prec_{BD})$, and
   (b) $w \not\models \alpha$ for all $w \in Min(Max(W, \prec_{BP}), \prec_{BD})$.
7. If β is either p or $\neg p$, then $M \models SBG_B(\alpha, \beta)$ iff
   (a) $M \models BEL_B(\beta \supset \alpha)$, and
   (b) $time(q) \leq time(p)$ for all $q \in atom(\alpha)$.
8. $M \models SBG_B(\alpha, \beta \wedge \gamma)$ iff there are β' and γ' such that
   (a) α is logically equivalent to $\beta' \wedge \gamma'$,
   (b) $atom(\alpha) = atom(\beta') \cup atom(\gamma')$, and
   (c) $M \models SBG_B(\beta', \beta) \wedge SBG_B(\gamma', \gamma)$.
9. $M \models SBG_B(\alpha, \neg(\beta \wedge \gamma))$ iff there are β' and γ' such that
   (a) α is logically equivalent to $\neg(\beta' \wedge \gamma')$,
   (b) $atom(\alpha) = atom(\beta') \cup atom(\gamma')$, and
   (c) $M \models SBG_B(\neg\beta', \neg\beta) \wedge SBG_B(\neg\gamma', \neg\gamma)$.
10. $M \models SBG_B(\alpha, \neg\neg\beta)$ iff $M \models SBG_B(\alpha, \beta)$.
11. $M \models GINT_B(\alpha)$ iff there is β such that $M \models \neg BEL_B(\alpha) \wedge SBG_B(\alpha, \beta) \wedge INT_B(\beta)$.
12. $M \models \phi \wedge \psi$ iff $M \models \phi$ and $M \models \psi$.
13. $M \models \neg\phi$ iff $M \not\models \phi$.

If $M \models \phi$, we call M a model of ϕ.
If $M \models \phi$ holds for all B-structures M, we write $\models \phi$.

12.1.3 A-Sentences and A-Structures

A-sentences are used to express facts about Agent A's mental attitudes. They are obtained by applying attitudinal operators for A, which are similar to those for B, to B-sentences.

Definition 12.5

1. If ϕ and ψ are B-sentences and $T \subset atom(\mathcal{L})$, then $BEL_A(\phi)$, $CHO_A(\phi)$, $P\text{-}PREF_A^T(\phi,\psi)$, $D\text{-}PREF_A^T(\phi,\psi)$, $INT_A(\phi)$, $SBG_A(\phi,\psi)$, and $GINT_A(\phi)$ are *A-sentences.*

2. If Φ and Ψ are *A-sentences*, then $\Phi \wedge \Psi$ and $\neg\Phi$ are *A-sentences.*

We denote A-sentences by $\Phi, \Psi, \ldots$.

Model structures for A-sentences are called *A-structures.*

Definition 12.6 $S = (\mathcal{M}, \prec_{AP}, \prec_{AD})$ is an *A-structure* iff

1. $\mathcal{M}$ is a nonempty set of B-structures, and

2. $\prec_{AP}$ and $\prec_{AD}$ are strict partial orders on B-structures which are bounded in $\mathcal{M}$.

An A-structure specifies Agent A's mental state: $\mathcal{M}$ specifies A's knowledge, $\prec_{AP}$ is A's plausibility order, and $\prec_{AD}$ is his desirability order. Again, we impose the condition that these orders are bounded, which is automatically satisfied when $atom(\mathcal{L})$ is finite.

Now we define the satisfaction relation for A-sentences. Motivations for the main part of the definition are explained in later sections.

Definition 12.7 Let $S = (\mathcal{M}, \prec_{AP}, \prec_{AD})$ be an A-structure.

1. $S \models BEL_A(\phi)$ iff $M \models \phi$ for all $M \in Max(\mathcal{M}, \prec_{AP})$.

2. $S \models CHO_A(\phi)$ iff $M \models \phi$ for all $M \in Max(Max(\mathcal{M}, \prec_{AP}), \prec_{AD})$.

3. $S \models P\text{-}PREF_A^T(\phi,\psi)$ iff $M_1 \prec_{AP} M_2$ for all pairs $M_1, M_2 \in \mathcal{M}$ such that

   (a) $M_1 \models \neg\phi \wedge \psi$,

   (b) $M_2 \models \phi \wedge \neg\psi$, and

   (c) $M_1 \models p$ iff $M_2 \models p$ for all $p \in T$.

4. $S \models D\text{-}PREF_A^T(\phi,\psi)$ iff $M_1 \prec_{AD} M_2$ for all pairs $M_1, M_2 \in Max(\mathcal{M}, \prec_{AP})$ such that

   (a) $M_1 \models \neg\phi \wedge \psi$,

(b) $M_2 \models \phi \wedge \neg\psi$, and

(c) $M_1 \models p$ iff $M_2 \models p$ for all $p \in T$.

5. $S \models INT_A(\phi)$ iff

   (a) $M \models \phi$ for all $M \in Max(Max(\mathcal{M}, \prec_{AP}), \prec_{AD})$, and

   (b) $M \not\models \phi$ for all $M \in Min(Max(\mathcal{M}, \prec_{AP}), \prec_{AD})$.

6. If ϕ or ψ contain attitudinal operators, $S \not\models SBG_A(\phi, \psi)$.

7. If ϕ contains no attitudinal operators and ψ is either p or $\neg p$, then $S \models SBG_A(\phi, \psi)$ iff

   (a) $S \models BEL_A(\psi \supset \phi)$, and

   (b) $time(q) \leq time(p)$ for all $q \in atom(\phi)$.

8. If ϕ, ψ and χ contain no attitudinal operators, then $S \models SBG_A(\phi, \psi \wedge \chi)$ iff there are ψ' and χ' such that

   (a) ϕ is logically equivalent to $\psi' \wedge \chi'$,

   (b) $atom(\phi) = atom(\psi') \cup atom(\chi')$, and

   (c) $M \models SBG_B(\psi', \psi) \wedge SBG_B(\chi', \chi)$.

9. If ϕ, ψ and χ contain no attitudinal operators, then $S \models SBG_A(\phi, \neg(\psi \wedge \chi))$ iff there are ψ' and χ' such that

   (a) ϕ is logically equivalent to $\neg(\psi' \wedge \chi')$,

   (b) $atom(\phi) = atom(\psi') \cup atom(\chi')$, and

   (c) $M \models SBG_B(\neg\psi', \neg\psi) \wedge SBG_B(\neg\chi', \neg\chi)$.

10. If ϕ contains no attitudinal operators, then $S \models SBG_A(\phi, \neg\neg\psi)$ iff $S \models SBG_A(\phi, \psi)$.

11. $S \models GINT_A(\phi)$ iff there is ψ such that $S \models \neg BEL_A(\phi) \wedge SBG_A(\phi, \psi) \wedge INT_A(\psi)$.

12. $S \models \Phi \wedge \Psi$ iff $S \models \Phi$ and $S \models \Psi$.

13. $S \models \neg\Phi$ iff $S \not\models \Phi$.

If $S \models \Phi$, we call S a model of Φ.
If $S \models \Phi$ holds for all A-structures S, we write $\models \Phi$.

To deal with the dynamics of mental attitudes, we make two assumptions about the dynamics of A-structures. First, we assume that the set M is monotonically reduced along the passage of time. It means that A's knowledge expands monotonically and A never gives up any piece of his knowledge. Second, we assume that the preference orders $\prec_{AP}$and $\prec_{AD}$do not change along the passage of time. We think this assumption is valid in daily situations, though preferences do change in a long time. Then we introduce a notion of monotonicity of attitudinal operators as follows:

Definition 12.8 An attitudinal operator *att* for A is *monotonic* iff
if $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models att(\phi)$ (or $att(\phi, \psi)$) and $\mathcal{M}'$ is a nonempty subset of $\mathcal{M}$, then $(\mathcal{M}', \prec_{AP}, \prec_{AD}) \models att(\phi)$ (or $att(\phi, \psi)$).

Most of attitudinal operators are nonmonotonic in our logic:

Theorem 12.1

1. *$P\text{-}PREF_A^T$ is monotonic, and*
2. *the other operators BEL_A, CHO_A, $D\text{-}PREF_A^T$, INT_A, SBG_A and $GINT_A$ are nonmonotonic.*

12.2 Belief

When an agent thinks about how the state of the world is, he considers only the most plausible possible worlds that are consistent with his knowledge. He believes sentences that are true in all these worlds. Therefore, we define Agent A's belief as follows: [1]

Definition 12.9 (Repeated) $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models BEL_A(\phi)$ iff
$M \models \phi$ for all $M \in Max(\mathcal{M}, \prec_{AP})$.

We can easily show that beliefs satisfy the following desirable properties: [2]

Theorem 12.2

1. *If $\models \phi$, then $\models BEL_A(\phi)$.*
2. $\models BEL_A(\phi) \supset \neg BEL_A(\neg\phi)$.
3. $\models BEL_A(\phi) \wedge BEL_A(\phi \supset \psi) \supset BEL_A(\psi)$.

[1] Hereafter, we omit definitions for Agent B, since they are similar to those for A.
[2] It is clear that B's attitudes satisfy similar properties to A's, though we do not mention them in this thesis.

Agent A's beliefs include all valid B-sentences, they are consistent, and they are closed under logical consequence. Note that we need boundedness of the plausibility order to obtain consistency of beliefs. In addition to these static properties, beliefs have dynamic properties. Nonmonotonicity is one of such properties, but more detailed analysis is presented as a theory of *belief revision* [14]. Satoh [37] examines the relationship between possible world model of belief with partial orders and the theory of belief revision, using a standard first-order language. Here, we try to get a similar result in our logic.

Gärdenfors [14] proposes a set of postulates for belief revision. A *belief set* K is a set of sentences that is closed under logical consequence. We denote by K_ϕ^* the revised set of sentences obtained from K by adding a sentence ϕ. We denote by K_ϕ^+ the set of all logical consequences of $K \cup \{\phi\}$. Then, Gärdenfors's postulates are stated as follows:

(K^*1) K_ϕ^* is a belief set.
(K^*2) $\phi \in K_\phi^*$.
(K^*3) $K_\phi^* \subset K_\phi^+$.
(K^*4) If $\neg\phi \notin K$, then $K_\phi^+ \subset K_\phi^*$.
(K^*5) K_ϕ^* is the set of all sentences if and only if $\models \neg\phi$.
(K^*6) If $\models \phi \equiv \psi$, then $K_\phi^* = K_\psi^*$.
(K^*7) $K_{\phi\wedge\psi}^* \subset (K_\phi^*)_\psi^+$.
(K^*8) If $\neg\psi \notin K_\phi^*$, then $(K_\phi^*)_\psi^+ \subset K_{\phi\wedge\psi}^*$.

Let $(\mathcal{M}, \prec_{AP}, \prec_{AD})$ be an A-structure. We define a belief set $B_\mathcal{M}$ by

$$B_\mathcal{M} = \{\chi \mid (\mathcal{M}, \prec_{AP}, \prec_{AD}) \models BEL_A(\chi)\}.$$

Since our belief sets are always consistent, we cannot define a revision function $*$ for contradictory inputs ϕ. We write $Mod(\phi)$ to mean the set of all models of ϕ. For ϕ such that $\mathcal{M} \cap Mod(\phi)$ is nonempty, we define a revision function $*$ by

$$(B_\mathcal{M})_\phi^* = B_{\mathcal{M}\cap Mod(\phi)}.$$

Then we get the following result:

Theorem 12.3 *For a belief set $K = B_\mathcal{M}$ and sentences ϕ and ψ such that $\mathcal{M} \cap Mod(\phi \wedge \psi)$ is nonempty, the revision function $*$ satisfies Gärdenfors's postulates (K^*1),(K^*2),(K^*3),(K^*5),(K^*6) and (K^*7).*

12.3 Choice

An agent plans and acts to make the most desirable possible worlds that are consistent with his belief true. He chooses the most desirable worlds to pursue, and as a result, he chooses all their consequences. Therefore, we say an agent chooses a sentence ϕ if ϕ is true in all the most desirable worlds that are consistent with his belief.

Definition 12.10 (Repeated) $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models CHO_A(\phi)$ iff $M \models \phi$ for all $M \in Max(Max(\mathcal{M}, \prec_{AP}), \prec_{AD})$.

We can easily show the following properties of choices:

Theorem 12.4

1. $\models BEL_A(\phi) \supset CHO_A(\phi)$.
2. $\models CHO_A(\phi) \supset \neg CHO_A(\neg\phi)$.
3. $\models CHO_A(\phi) \wedge CHO_A(\phi \supset \psi) \supset CHO_A(\psi)$.

Agent A's choices include all his beliefs, they are consistent, and they are closed under logical consequence.

An agent's choices are not always desirable for him. For example, an agent who chooses his action of eating lunch at a restaurant also chooses his spending money on the bill, which is probably not desirable for him:

$$\begin{aligned} \models\ & CHO_A(eatA_1) \wedge \\ & BEL_A(eatA_1 \supset spendMoneyA_2) \\ \supset\ & CHO_A(spendMoneyA_2). \end{aligned}$$

Choices determine the output of an agent, that is, actions. Although performance of actions is outside our logic, we make several informal assumptions about the relationship between A-structures and Agent A's performance of actions. Our first assumption says that Agent A performs an action when he chooses it:

($A1$) If $S \models CHO_A(actA_i)$ and i is the time of S, then A performs act.

The converse of ($A1$) does not hold because of well-known Buridan cases [5]. Consider that there are two actions act and act' that are equally desirable for Agent A to do. In this case, some of the most desirable worlds include A's performance of act and the others of them include his performance of act'. According to our definition, Agent A chooses neither one of these actions, but in reality, he after all performs one of them. Therefore, we adopt the following weaker assumption:

($A2$) If A performs act, then $S \models \neg CHO_A(\neg actA_i)$.

An agent's actions do not necessarily make all his choices true. An agent never chooses implausible sentences, but he may choose contingent sentences. For example, an agent may choose its being fine the next day, though he has nothing to do to make it true. We might restrict the notion of choice to make its relationship with actions closer if needed.

12.4 Preference

In this section, we investigate a notion of preference between sentences. Like preference for model structures, there are two types of preference, that is, preference about plausibility and preference about desirability. As mentioned in Section 11.1, this notion is somewhat ambiguous and can be reduced to preference ordering on model structures in two different ways. The first possible reduction of a preference for a sentence ϕ to another sentence ψ is that models of ϕ are collectively or averagely preferred to models of ψ. This means that ϕ is materially preferred to ψ with respect to the current situation. For example, a sentence

"It is more likely to rain than to snow tomorrow."

expresses a preference of this type, that the probability of its raining the next day is higher than the probability of its snowing. Such a preference is studied in connection with *conditionals* [46, 51]. We can partly deal with this type of preference using the most preferred structures. That is, Agent A plausibly prefers ϕ to ψ if $\neg BEL_A(\neg\phi) \wedge BEL_A(\neg\psi)$ holds. Similarly, Agent A desirably prefers ϕ to ψ if $\neg CHO_A(\neg\phi) \wedge CHO_A(\neg\psi)$ holds.

The second possible reduction of the preference for ϕ to ψ is that each model of ϕ is preferred to each model of ψ. This means a general fact, that ϕ is preferred to ψ under all conditions. For example, a sentence

"In summer, it is more likely to rain than to snow."

expresses a preference of this type, that for every day in summer the probability of its raining is higher than the probability of its snowing. An agent learns and uses many such general preferences in his daily life. In this thesis, we deal with preferences of this second type. We said that we compare each model of ϕ and each model of ψ, but this statement is not correct because models of $\phi \wedge \psi$ must be excluded from the comparison. In fact, we compare each model of $\phi \wedge \neg\psi$ and each model of $\neg\phi \wedge \psi$. Moreover, we must decide whether the parts of the world other than ϕ and ψ must be fixed or may vary through the comparison. We take a general approach and use a parameter $T \subset atom(\mathcal{L})$ to specify which atoms must be fixed. Now we define plausibility preference $P\text{-}PREF_A^T$ and desirability preference $D\text{-}PREF_A^T$ as follows:

Definition 12.11 (Repeated)

1. $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models P\text{-}PREF_A^T(\phi, \psi)$ iff
   $M_1 \prec_{AP} M_2$ for all pairs $M_1, M_2 \in \mathcal{M}$ such that
   (a) $M_1 \models \neg\phi \wedge \psi$,
   (b) $M_2 \models \phi \wedge \neg\psi$, and
   (c) $M_1 \models p$ iff $M_2 \models p$ for all $p \in T$.

2. $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models D\text{-}PREF_A^T(\phi, \psi)$ iff
$M_1 \prec_{AD} M_2$ for all pairs $M_1, M_2 \in Max(\mathcal{M}, \prec_{AP})$ such that

   (a) $M_1 \models \neg\phi \wedge \psi$,

   (b) $M_2 \models \phi \wedge \neg\psi$, and

   (c) $M_1 \models p$ iff $M_2 \models p$ for all $p \in T$.

It is convenient if we have a notation for expressing conditional preferences. To express a preference for ϕ to ψ under a condition χ, we introduce the following abbreviations:

$$P\text{-}PREF_A^T(\phi, \psi \mid \chi) = P\text{-}PREF_A^T(\chi \supset \phi, \chi \supset \psi).$$
$$D\text{-}PREF_A^T(\phi, \psi \mid \chi) = D\text{-}PREF_A^T(\chi \supset \phi, \chi \supset \psi).$$

First, we list general properties that hold for all T and for both $P\text{-}PREF_A^T$ and $D\text{-}PREF_A^T$ (written $X\text{-}PREF_A^T$, for short).

Theorem 12.5

1. *If* $\models \phi \supset \psi$, *then* $\models X\text{-}PREF_A^T(\phi, \psi) \wedge X\text{-}PREF_A^T(\psi, \phi)$.
In particular,

   (a) $\models X\text{-}PREF_A^T(\phi, \phi)$.

   (b) If ϕ *or* ψ *is either a tautology or a falsity,* $\models X\text{-}PREF_A^T(\phi, \psi)$.

2. *If* $\models (\phi \equiv \phi') \wedge (\psi \equiv \psi')$, *then* $\models X\text{-}PREF_A^T(\phi, \psi) \equiv X\text{-}PREF_A^T(\phi', \psi')$.

3. $\models X\text{-}PREF_A^T(\phi, \psi) \equiv X\text{-}PREF_A^T(\neg\psi, \neg\phi)$.

4. $\models X\text{-}PREF_A^T(\phi, \psi \mid \chi) \equiv X\text{-}PREF_A^T(\phi \wedge \chi, \psi \wedge \chi)$.

5. $\models X\text{-}PREF_A^T(\phi, \psi) \supset X\text{-}PREF_A^T(\phi, \psi \mid \chi)$.

6. *If* $p \in T$,
$\models X\text{-}PREF_A^T(\phi, \psi \mid p) \wedge X\text{-}PREF_A^T(\phi, \psi \mid \neg p) \supset X\text{-}PREF_A^T(\phi, \psi)$.

7. *If* $T \subset T'$, $\models X\text{-}PREF_A^T(\phi, \psi) \supset X\text{-}PREF_A^{T'}(\phi, \psi)$.

Item 1 of the theorem says that for two sentences one of which is a logical consequence of the other, preference relations always hold. 3 says that preferring ϕ to ψ is equal to preferring $\neg\psi$ to $\neg\phi$. 4 deals with a preference under a condition χ. From 5, we see that if ϕ is preferred to ψ, it is preferred under all conditions. On the other hand, ϕ is preferred to ψ if it is preferred under both conditions p and $\neg p$ for some atom p as is shown in 6. 7 says that if a parameter set T is the smaller, the preference $X\text{-}PREF_A^T$ is the stronger. Note that transitivity of preferences

$$X\text{-}PREF_A^T(\phi, \psi) \wedge X\text{-}PREF_A^T(\psi, \chi) \supset X\text{-}PREF_A^T(\phi, \chi)$$

does not hold. To obtain a counterexample, let $\psi = \phi \wedge \chi$.

Although we allow T to be any set of atoms, we mainly use three patterns of T, for which we introduce the following notations (Y denotes either A or B):

(1) $X\text{-}PREF_Y^{\emptyset}(\phi, \psi) = X\text{-}PREF_Y^{T}(\phi, \psi)$ where T is an empty set.
(2) $X\text{-}PREF_Y^{i,j}(\phi, \psi) = X\text{-}PREF_Y^{T}(\phi, \psi)$
where $T = \{p \in atom(\mathcal{L}) \mid i < time(p) < j\}$.
(3) $X\text{-}PREF_Y^{eq}(\phi, \psi) = X\text{-}PREF_Y^{T}(\phi, \psi)$
where $T = atom(\mathcal{L}) \setminus (atom(\phi) \cup atom(\psi))$.

(1) expresses the strongest preference, that ϕ is preferred to ψ unconditionally. This type of sentence is widely used to express an agent's default preferences about plausibility and desirability. For example, a sentence $D\text{-}PREF_A^{\emptyset}(\neg hungry, hungry)$ means that Agent A prefers to be not hungry even if he must spend much money and time to satisfy his hunger. The strongest preference (1) satisfies strong and interesting properties. For example, it satisfies the following restricted form of transitivity:

Theorem 12.6

1. $\models X\text{-}PREF_A^{\emptyset}(\phi, \psi) \wedge X\text{-}PREF_A^{\emptyset}(\psi, \chi) \wedge \neg BEL_A(\phi \wedge \chi \supset \psi) \supset X\text{-}PREF_A^{\emptyset}(\phi, \chi \mid \psi)$.
2. $\models X\text{-}PREF_A^{\emptyset}(\phi, \psi) \wedge X\text{-}PREF_A^{\emptyset}(\psi, \chi) \wedge \neg BEL_A(\psi \supset \phi \vee \chi) \supset X\text{-}PREF_A^{\emptyset}(\phi, \chi \mid \neg\psi)$.

(2) is used to express a kind of *frame axioms* [29], which are representations of persistency of properties. A sentence $P\text{-}PREF_A^{1,3}(atHome_3, \neg atHome_3 \mid atHome_1)$ expresses the fact that if Agent A is at home at time 1, he is more likely to be at home at time 3 than to be not at home if states of the world between time 1 and time 3 (namely, at time 2) are equal.

(3) expresses the weakest preference, that ϕ is preferred to ψ if all else are equal. If ϕ and ψ cause different effects on other parts of the world, they are not compared. For example, a sentence $D\text{-}PREF_A^{eq}(\neg hungry, hungry)$ means that Agent A prefers to be not hungry if the other conditions are equal. In this case, he may not want to spend money and time to satisfy his hunger.

Preference between sentences has a close relationship with other attitudes. In particular, for the strongest type of preferences where $T = \emptyset$, we get the following interesting results:

Theorem 12.7

1. $\models P\text{-}PREF_A^{\emptyset}(\phi, \neg\phi) \wedge \neg BEL_A(\neg\phi) \supset BEL_A(\phi)$.
2. $\models D\text{-}PREF_A^{\emptyset}(\phi, \neg\phi) \wedge \neg BEL_A(\neg\phi) \supset CHO_A(\phi)$.
3. $\models P\text{-}PREF_A^{\emptyset}(\phi, \psi) \wedge \neg BEL_A(\phi \supset \psi) \supset BEL_A(\psi \supset \phi)$.
4. $\models D\text{-}PREF_A^{\emptyset}(\phi, \psi) \wedge \neg BEL_A(\phi \supset \psi) \supset CHO_A(\psi \supset \phi)$.

5. $\models P\text{-}PREF_A^{\emptyset}(\phi, \neg\phi \mid \psi) \wedge \neg BEL_A(\psi \supset \neg\phi) \supset BEL_A(\psi \supset \phi)$.

6. $\models D\text{-}PREF_A^{\emptyset}(\phi, \neg\phi \mid \psi) \wedge \neg BEL_A(\psi \supset \neg\phi) \supset CHO_A(\psi \supset \phi)$.

Item 1, 2 of the theorem show that preference of the form $X\text{-}PREF_A^{\emptyset}(\phi, \neg\phi)$ correspond to a notion of *defaults*: if ϕ is consistent with his belief, Agent A believes (chooses) ϕ. These results are generalized in two ways, which are shown in Items 3, 4 and Items 5, 6. If Agent A prefers ϕ to ψ, he normally believes (chooses) $\psi \supset \phi$. Preferences of the form $X\text{-}PREF_A^{\emptyset}(\phi, \neg\phi \mid \psi)$ correspond to *conditional defaults*: Agent A normally believes (chooses) $\psi \supset \phi$.

Before ending this section, we introduce a notion of the *order represented by* a sentence, which is used in the next section.

Definition 12.12 The *order* $\prec X\text{-}PREF_A^T(\phi, \psi)$ *represented by* a sentence $X\text{-}PREF_A^T(\phi, \psi)$ (X is either P or D) is a strict partial order on B-structure defined as follows: $M_1 \prec X\text{-}PREF_A^T(\phi, \psi) M_2$ iff

1. $M_1 \models \neg\phi \wedge \psi$,

2. $M_2 \models \phi \wedge \neg\psi$, and

3. $M_1 \models p$ iff $M_2 \models p$ for all $p \in T$.

With this notion, we can rephrase the satisfaction relation for $P\text{-}PREF_A^T$ and $D\text{-}PREF_A^T$ as follows:

Theorem 12.8

1. $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models P\text{-}PREF_A^T(\phi, \psi)$ *iff*
$\prec P\text{-}PREF_A^T(\phi, \psi) \cap (\mathcal{M} \times \mathcal{M}) \subset \prec_{AP}$.

2. $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models D\text{-}PREF_A^T(\phi, \psi)$ *iff*
$\prec D\text{-}PREF_A^T(\phi, \psi) \cap (Max(\mathcal{M}, \prec_{AP}) \times Max(\mathcal{M}, precap)) \subset \prec_{AD}$.

12.5 Structure Specification Lists

Agent A's mental state is represented by an A-structure $(\mathcal{M}, \prec_{AP}, \prec_{AD})$, which consists of a knowledge state $\mathcal{M}$, a plausibility order $\prec_{AP}$, and a desirability order $\prec_{AD}$. In this section, we introduce constructs of sentences called *structure specification lists*, which are used to specify these constructs of A-structures. For a set $\mathcal{X}$ of sentences, we define $Mod(\mathcal{X}) = \bigcap_{\phi \in \mathcal{X}} Mod(\phi)$. We say $\mathcal{X}$ is consistent if $Mod(\mathcal{X})$ is nonempty.

Definition 12.13 $U = (\mathcal{K}, \mathcal{P}, \mathcal{D})$ is a *structure specification list* iff

1. $\mathcal{K}$ is a consistent set of B-sentences.

2. $\mathcal{P}$ is a set of sentences of the form $P\text{-}PREF_A^T(\phi, \psi)$ which is well-ordered by a relation "*precedes*". (That is, for all nonempty $\mathcal{P}' \subset \mathcal{P}$, there is a $P \in \mathcal{P}'$ that *precedes* all the other elements of $\mathcal{P}'$.)

3. $\mathcal{D}$ is a set of sentences of the form $D\text{-}PREF_A^T(\phi, \psi)$ which is well-ordered by the relation "*precedes*".

A knowledge state $\mathcal{M}$ is specified by a consistent set $\mathcal{K}$ of B-sentences, which corresponds to an axiom system possessed by Agent A. We simply let $\mathcal{M} = Mod(\mathcal{K})$.

A plausibility order $\prec_{AP}$is specified by an ordered set $\mathcal{P}$ of sentences of the form $P\text{-}PREF_A^T(\phi, \psi)$, and a desirability order $\prec_{AD}$is specified by an ordered set $\mathcal{D}$ of sentences of the form $D\text{-}PREF_A^T(\phi, \psi)$. That is, we construct preference orders from Agent A's preferences between sentences. An agent generally has many preferences, to which he gives priority ranking. His preferences may conflict with each other, and such conflicts are solved according to priority.

There are two typical sources of conflicts. First, general preferences are often overridden by more specific preferences. For example, a preference (a conditional default) that animals do not normally fly $P\text{-}PREF_A^{eq}(\neg fly, fly \mid animal)$ is overridden by a more specific preference that birds normally fly $P\text{-}PREF_A^{eq}(fly, \neg fly \mid bird)$.

Second, an agent usually has conflicting desires. Consider that choosing a restaurant to eat lunch, Agent A has both of the following preferences:

(1) $D\text{-}PREF_A^{\emptyset}(\neg atCrowded, atCrowded)$
(A prefers not to eat at a crowded restaurant.)
(2) $D\text{-}PREF_A^{\emptyset}(eatSushi, eatSoba)$
(A prefers to eat sushi rather than to eat soba.)

Then, which restaurant does he prefer, a crowded sushi bar or a noodle shop that is not crowded? The answer depends on his priority for these preferences: if (1) has higher priority than (2), he prefers the noodle shop that is not crowded. If (2) has higher priority than (1), he prefers the crowded sushi bar.

Before we give a method of constructing preference orders from sets of sentences, we clarify a notion of conflict among preferences. For a binary relation R, we write R^+ to mean the transitive closure of R.

Definition 12.14 A set $\mathcal{X}$ of sentences of the form $X\text{-}PREF_A^T(\phi, \psi)$ (X is either P or D) is *compatible* iff $(\bigcup_{X \in \mathcal{X}} \prec X)^+$ is a strict partial order.

Remember that $\prec X$ is the order represented by a sentence X. For example, both of the following sets are compatible:

$$\{P\text{-}PREF_A^{\emptyset}(\phi, \psi \mid \chi), P\text{-}PREF_A^{\emptyset}(\phi', \psi' \mid \neg\chi)\},$$
$$\{D\text{-}PREF_A^{eq}(p, q), D\text{-}PREF_A^{eq}(r, s)\}$$
where p, q, r and s are distinct atoms of $\mathcal{L}$.

Now we define the order $\prec_{\mathcal{X}}$ on B-structures that is represented by an well-ordered set of sentences $\mathcal{X}$. Note that if $\mathcal{X}$ is compatible, we can simply put $\prec_{\mathcal{X}} = (\bigcup_{X \in \mathcal{X}} \prec X)^+$. In general, however, we need to remove parts of preferences that cause conflicts, according to priority. For a binary relation R, we write R^{-1} to mean the inverse relation of R.

Definition 12.15 Let $\mathcal{X}$ be a set of sentences of the form $X\text{-}PREF_A^T(\phi, \psi)$ which is well-ordered by a relation *precedes*. Then, we define

$$\prec_{\mathcal{X}} = (\bigcup_{X \in \mathcal{X}} \overline{\prec X})^+$$

where $\overline{\prec X}$ is defined by transfinite induction as follows:

$$\overline{\prec X} = \prec X - ((\prec X \cup \bigcup_{X' \, precedes \, X} \overline{\prec X'})^+)^{-1}.$$

For every $X \in \mathcal{X}$, we remove elements of $\prec X$ that cause conflicts when they are linked with other preferences that have higher priority. We say $\mathcal{X}'$ is an *initial segment* of $\mathcal{X}$ if $\mathcal{X}' \subset \mathcal{X}$ and every element of $\mathcal{X}'$ precedes every element of $\mathcal{X} - \mathcal{X}'$. We have the following desirable results:

Theorem 12.9

1. $\prec_{\mathcal{X}}$ *is a strict partial order.*
2. $\prec_{\mathcal{X}} \subset (\bigcup_{X \in \mathcal{X}} \prec X)^+$.
3. *If* $\mathcal{X}$ *is compatible,* $\prec_{\mathcal{X}} = (\bigcup_{X \in \mathcal{X}} \prec X)^+$.
4. *If* $\mathcal{X}'$ *is an initial segment of* $\mathcal{X}$, $\prec_{\mathcal{X}'} \subset \prec_{\mathcal{X}}$.

It is useful if we can transform a set of sentences into a compatible set of sentences that represents the same order. We show that for a finite set of sentences of the form $X\text{-}PREF_A^{\emptyset}(\phi, \psi)$ such transformation is possible. In order to represent an well-ordered set of sentences, we often use a list notation $\langle X, X', \ldots, X'', \ldots \rangle$ which lists its elements in precedence order.

Let $\mathcal{X} = \langle X_1, \ldots, X_n \rangle$ where $X_i = X\text{-}PREF_A^{\emptyset}(\phi_i, \psi_i)$ for all $i = 1, \ldots, n$. For all $i = 1, \ldots, n$, we define

$$\begin{aligned} C(X_i) \;=\; & \{X\text{-}PREF_A^{\emptyset}(\phi_i, \psi_i \mid C_1 \wedge \cdots \wedge C_{i-1}) \mid \\ & C_j \text{ is either } \phi_j \vee \neg\psi_j \text{ or } \neg\phi_j \vee \psi_j \text{ for all } j = 1, \ldots, i-1\} \end{aligned}$$

and then define

$$C(\mathcal{X}) = \bigcup_{i=1}^{n} C(X_i).$$

We can consider the precedes relation on $C(\mathcal{X})$ to be any well-order, since it is not essential in compatible sets. Then we can easily show that $C(\mathcal{X})$ is compatible and $\prec_{C(\mathcal{X})}=\prec_{\mathcal{X}}$.

We give an example of transformation as follows:

$$\begin{aligned}
\mathcal{X} &= \langle D\text{-}PREF_A^{\emptyset}(\neg atCrowded, atCrowded), \\
&\quad D\text{-}PREF_A^{\emptyset}(eatSushi, eatSoba) \\
&\quad \rangle. \\
C(\mathcal{X}) &= \langle D\text{-}PREF_A^{\emptyset}(\neg atCrowded, atCrowded), \\
&\quad D\text{-}PREF_A^{\emptyset}(eatSushi, eatSoba \mid \neg atCrowded), \\
&\quad D\text{-}PREF_A^{\emptyset}(eatSushi, eatSoba \mid atCrowded) \\
&\quad \rangle.
\end{aligned}$$

Let us return to structure specification lists. A structure specification list $U = (\mathcal{K}, \mathcal{P}, \mathcal{D})$ specifies an A-structure $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}})$ if both $\prec_{\mathcal{P}}$ and $\prec_{\mathcal{D}}$ are bounded in $Mod(\mathcal{K})$. Therefore, we can regard a structure specification list U as a model of Agent A. An input to U is a new piece of knowledge expressed by a B-sentence ϕ. Then, the new state of Agent A is represented by $U' = (\mathcal{K} \cup \{\phi\}, \mathcal{P}, \mathcal{D})$. The output of U obeys assumptions presented in Section 12.3. A-structures specified by structure specification lists satisfy the following properties:

Theorem 12.10 *Let $(\mathcal{K}, \mathcal{P}, \mathcal{D})$ be a structure specification list that specifies an A-structure $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}})$.*

1. *For all $\phi \in \mathcal{K}$, $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models BEL_A(\phi)$.*
2. *If $\mathcal{P}'$ is a compatible initial segment of $\mathcal{P}$, then for all $P \in \mathcal{P}'$, $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec \mathcal{D}) \models P$.*
3. *If $\mathcal{D}'$ is a compatible initial segment of $\mathcal{D}$, then for all $D \in \mathcal{D}'$, $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec \mathcal{D}) \models D$.*
4. *If $\mathcal{P}'$ is an initial segment of $\mathcal{P}$ and $(Mod(\mathcal{K}), \prec_{\mathcal{P}'}, \prec_{\mathcal{D}}) \models BEL_A(\phi)$, then $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models BEL_A(\phi)$.*
5. *If $\mathcal{D}'$ is an initial segment of $\mathcal{D}$ and $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}'}) \models CHO_A(\phi)$, then $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models CHO_A(\phi)$.*

Chapter 13

Intention and Generalized Intention

13.1 Introduction

In recent years, it has become increasingly obvious that intention, a kind of mental attitudes of an agent, plays an important role in communications in multi-agent environment [11, 27, 40]. An agent recognizes his companion's intentions to predict his actions and to behave cooperatively. Conversely, an agent indicates his intentions to his companion by actions. Natural language is a sophisticated tool for expressing intentions. For example, the following sentences express the same intention of the speaker, that is, an intention of knowing where the bus starts.

"I want to know where the bus starts."
"Please tell me where the bus starts."
"Do you know where the bus starts?"

The relationship between natural language sentences and intentions that they express is studied in a theory of speech acts [40, 41].

In this chapter, we give a logical formulation of intention by using preference of an agent. First, we need to clarify a notion of intention that we try to formulate, since intention is a notoriously ambiguous notion. Intuitively speaking, an intention of an agent is a sentence ϕ such that (1) the agent is going to achieve ϕ, and (2) ϕ is desirable for him. With what we have defined already, Condition (1) is paraphrased into a condition that ϕ is a choice of the agent. Therefore, we consider intention to be a kind of choice, that is, desirable choice, though a notion of desirable sentences has not formally defined yet. Note that all choices are not intentions. For example, consider an agent who believes a sentence

$$eatSushi_2 \supset spendMoney_3 \wedge goSushiBar_1$$

which expresses that if he eats sushi at a sushi bar, he spends money after eating and goes to that bar before eating. Assume that he intends $eatSushi_2$. Since

intentions are choices and choices are closed under logical consequence, we see that both $spendMoney_3$ and $goSushiBar_1$ are his choices. $spendMoney_3$ is obviously not an intention, since it is not desirable for the agent. We think that $goSushiBar_1$ is also not an intention, since it is not desirable for itself. (Consider a situation where the agent has no money and he is unable to eat sushi.) But unlike $spendMoney_3$, we can consider $goSushiBar_1$ to be desired by the agent as a way to achieve $eatSushi_2$, and thus we consider it to be an intention in a general sense. In fact, it corresponds to a notion of subgoal used in planning theory. Although we mainly deal with intention (INT) which is desirable for itself in this chapter, in the last section of the chapter we formulate notions of subgoal (SBG) and generalized intention ($GINT$).

The following is the main properties of intentions ever proposed in the literature [5, 33, 36]:

1. Intentions are consistent with each other and with beliefs.
2. Intentions are not believed to hold already.
3. Intentions are *not* closed under logical consequence.
4. Intentions have *persistency*.
5. Intentions are closely related to preferences about desirability.

Unlike simple desires, intentions are sentences an agent is going to achieve, and thus they must be consistent. If an agent believes that a sentence ϕ is satisfied already, he never intends to make ϕ true, since such an intention is unnecessary. Intentions are generally not closed under logical consequence; an agent does not need to intend all consequences of his intentions. For example, an agent who intends ϕ (e.g., going to a library) does not need to intend $\phi \vee \psi$ (e.g., going to a zoo). Intentions are nonmonotonic, but they have certain persistency. Once an agent adopts an intention, he never drops it without special reasons. Intentions are related to preferences in various ways. If two intentions of an agent become inconsistent, the agent gives up the intention which he does not prefer to the other intention. If an agent knows two plans for achieving an intention and he prefers one plan to the other, he usually intends to perform the preferred plan. Conversely, if an agent intends to perform a plan, we can infer under certain conditions that he intends the result of the plan and that he prefers that plan to the other plans.

The first attempt to formalize intention is made by Cohen and Levesque [10]. They used persistency to characterize intentions. According to their definition, a choice [1] is persistent if it will not be dropped until the agent thinks it has been satisfied or he thinks it will never be true. Then they identified intention with a special kind of persistent choice. Although defined intentions have strong and interesting consequences, there are two problems. First, persistent choices are essentially closed under tautological consequence. For example, if ϕ is a persistent choice, then so is

[1] They use a term "goal" to mean what we call "choice".

$\phi \vee \psi$ for arbitrary ψ. Second, their definition of persistency is too strong, since an agent may drop his intentions for various reasons. For instance, if an agent comes to know that his intentions ϕ and ψ are mutually exclusive, he must give up at least one of these intentions, while each of them may be still achievable.

Konolige and Pollack [25] took another approach, a representationalist approach to intention. They represent intentions directly in a cognitive structure of an agent. Restricting admissible structures, they showed that intentions can satisfy desirable basic properties without suffering from the consequential closure. However, the dynamics of intention must be provided from outside the model, therefore we cannot at all examine such properties as persistency in their formalism.

Moreover, neither of these theories considers the relationship between intentions and preferences of an agent. We think it is a severe limitation of them particularly when we deal with interactions among intentions and reasoning about plans.

In this thesis, we define intention in terms of desirability preference. Defined intentions satisfy all of the five properties mentioned above. In particular, a new preference-based account for the consequential closure problem and persistency is given. Bratman [5] argues that intention is not reducible to a combination of other mental attitudes like belief and desire, because intention concerns an agent's commitment to future actions. An agent has incomplete reasoning capability and limited resources, and thus his actions do not necessarily agree with his preference all the time. Although this argument is correct, it is also true that an agent's adoption and abandonment of intentions are determined mainly by his preferences. Therefore, we think our formulation properly model most aspects of intention of a rational agent.

13.2 Intention and Preference

To define intention as desirable choice, we need to formulate a notion that a sentence ϕ is desirable for Agent A. This notion is ambiguous, and it is hard to choose the best definition from possible ones, though some of them are clearly inadequate; for example, a condition $D\text{-}PREF^{\emptyset}_{A}(\phi, \neg\phi)$ is clearly too strong for ϕ's being desirable for A. Therefore, we define this notion simply by a minimal requirement that ϕ is *not bad* for Agent A. That is, ϕ is desirable for A if ϕ is satisfied in no minimally preferred structure. This requirement is reasonable, because if there exists a model of ϕ that is minimally preferred, attempting to satisfy ϕ may result in that minimally preferred model, that is, may not make the things better, and thus there is no reason for Agent A to pursue ϕ. Now intention is defined as follows:

Definition 13.1 (Repeated) $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models INT_A(\phi)$ iff

1. $M \models \phi$ for all $M \in Max(Max(\mathcal{M}, \prec_{AP}), \prec_{AD})$, and
2. $M \not\models \phi$ for all $M \in Min(Max(\mathcal{M}, \prec_{AP}), \prec_{AD})$.

We can easily show the following properties of intentions:

Theorem 13.1

1. $\models INT_A(\phi) \supset CHO_A(\phi)$.
2. $\models INT_A(\phi) \supset \neg BEL_A(\phi) \wedge \neg BEL_A(\neg\phi)$.
3. $\models INT_A(\phi) \wedge BEL_A(\phi \equiv \psi) \supset INT_A(\psi)$.
4. $\models INT_A(\phi) \wedge INT_A(\psi) \supset INT_A(\phi \wedge \psi)$.
5. $\models INT_A(\phi) \wedge INT_A(\psi) \supset INT_A(\phi \vee \psi)$.

An agent does not intend sentences which he believes to be satisfied already. Intentions are consistent with each other and with beliefs, and they are closed under conjunction and disjunction. The converse of 4 and 5 does not hold in general. Intending a conjunction does not imply intending its conjuncts. However, if one conjunct is preferred to the other, we can conclude that at least the preferred conjunct is intended:

Theorem 13.2

1. $\models INT_A(\phi \wedge \psi) \wedge D\text{-}PREF_A^{\emptyset}(\phi, \psi) \wedge \neg BEL_A(\psi \supset \phi) \supset INT_A(\phi)$.
2. $\models INT_A(\phi \vee \psi) \wedge D\text{-}PREF_A^{\emptyset}(\phi, \psi) \wedge \neg BEL_A(\phi \supset \psi) \supset INT_A(\phi)$.

To demonstrate that our definition agrees with our intuitions, let us take an example. Consider that Agent A goes to a bookstore intending to buy a paperback and also intending to buy a magazine, because he likes to buy them. This situation is described by the following structure specification list:

$$\begin{aligned}
&\mathcal{K} = \{\,\} \\
&\mathcal{P} = \langle\,\rangle \\
&\mathcal{D} = \langle\ D\text{-}PREF_A^{eq}(buyPaperback, \neg buyPaperback), \\
&\qquad D\text{-}PREF_A^{eq}(buyMagazine, \neg buyMagazine) \\
&\quad \rangle\,.
\end{aligned}$$

Then we have the following:

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \ \models\ INT_A(buyPaperback) \wedge INT_A(buyMagazine).$$

Consider a sentence $buyPaperback \equiv buyMagazine$ which means buying both or neither. Although this sentence is a tautological consequence of an intention $buyPaperback \wedge buyMagazine$ and hence it is a choice, it is *not* intended, because $\neg buyPaperback \wedge \neg buyMagazine$ is satisfied in minimally preferred models. This shows that intentions are not closed under tautological consequence in our logic:

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \ \models\ INT_A(buyPaperback \wedge buyMagazine) \wedge \neg INT_A(buyPaperback \equiv buyMagazine).$$

At a bookstore, Agent A happens to know that he has not enough money to buy both of them. If he has no preference between buying the paperback and buying the magazine, he drops intentions to buy each of them, and now only intends to buy one of them:

$$\begin{aligned}(Mod(\mathcal{K}'), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models\ & INT_A(buyPaperback \vee buyMagazine) \wedge \\ & \neg INT_A(buyPaperback) \wedge \\ & \neg INT_A(buyMagazine)\end{aligned}$$

where the new state of knowledge $\mathcal{K}'$ is given as follows:

$$\mathcal{K}' = \{\, \neg buyPaperback \vee \neg buyMagazine \}.$$

On the other hand, if he prefers the paperback to the magazine, he continues intending to buy the paperback, while he gives up the intention to buy the magazine:

$$\begin{aligned}(Mod(\mathcal{K}'), \prec_{\mathcal{P}}, \prec_{\mathcal{D}'}) \models\ & INT_A(buyPaperback \vee buyMagazine) \wedge \\ & INT_A(buyPaperback) \wedge \\ & \neg INT_A(buyMagazine)\end{aligned}$$

where

$$\begin{aligned}\mathcal{D}' = \langle\ & D\text{-}PREF_A^{eq}(buyPaperback, \neg buyPaperback), \\ & D\text{-}PREF_A^{eq}(buyMagazine, \neg buyMagazine), \\ & D\text{-}PREF_A^{eq}(buyPaperback, buyMagazine) \\ \rangle\ &.\end{aligned}$$

We have shown that intentions defined in our logic satisfy good properties. They are consistent with each other and with beliefs. They are not believed to hold already. They are not closed under logical consequence. They are closely related to preferences, as is demonstrated by Theorem 13.2 and the last example. (We give further examples in the next chapter.) Moreover, we show that they have a kind of persistency in the next section.

13.3 Persistency of Intentions

Intentions have *persistency*. Once an agent adopts an intention, he never drops it without special reasons. Bratman [5] argues that this property of intentions is essential to the practical reasoning capability of an agent, since the stability of intentions enables an agent to coordinate his current activities in accordance with his intentions about the future. Persistency mainly stems from resource limitations of an agent, who refuses reconsideration of intentions. Cohen and Levesque [10] treat persistency as a characteristic of intentions that distinguishes them from other simple choices. According to their definition, an intention is a special kind of choice

which will never be dropped until an agent thinks it has been satisfied or he thinks it is impossible to achieve.

However, this formulation of persistency is obviously too strong. We already have a counterexample, the bookstore example presented in the last section. Agent A drops the intention to buy the magazine although he does not think either that he has bought the magazine or that he cannot buy the magazine. It only conflicts with another intention, the intention to buy the paperback. Cohen and Levesque also introduce in [10] a notion of *relativized* persistency, persistency as long as some pre-specified condition is consistent. But, it is not clear whether such conditions are exhaustively expressible in real situations. After all, persistency of intentions is a prototypical phenomenon and we cannot completely specify the condition of when to give up an intention.

In our view, there is another type of persistency which comes from invariability of the preference order of an agent. An agent prefers his intentions, and if his preference order does not change, the agent will continue to have those intentions. Of course, it is not always the case. As illustrated by the example, an intention is dropped when it becomes inconsistent with other intentions to which it is not preferred. But even in such cases, if the agent later thinks its rivals are unachievable, the intention will be recovered. In our example, if the paperback is sold out and Agent A comes to know that fact, he again adopts the intention to buy the magazine. [2] This suggests that we can formalize a weaker form of persistency of intentions, but before we can proceed, some preparations are required. Here, we only deal with monotonic growth of belief, that is, we assume that the plausibility order is an empty order $\emptyset$.

Definition 13.2 $\mathcal{M}'$ is a *partial restriction of* $\mathcal{M}$ *with respect to* ϕ iff

1. $\mathcal{M}' \subset \mathcal{M}$,
2. $\mathcal{M}' \setminus Mod(\phi) = \mathcal{M} \setminus Mod(\phi)$, and
3. $\mathcal{M}' \cap Mod(\phi)$ is nonempty.

A partial restriction with respect to ϕ corresponds to Agent A's adopting a new belief $\phi \supset \psi$ for some ψ. It restricts models of ϕ *partially*, in the sense that it never makes ϕ inconsistent. The following theorem says that if Agent A drops an intention ϕ by adopting a belief about ϕ, it is possible that he will recover that intention later. [3]

Theorem 13.3 *If* $(\mathcal{M}, \emptyset, \prec_{AD}) \models INT_A(\phi)$, *then for every partial restriction* $\mathcal{M}'$ *of* $\mathcal{M}$ *with respect to* ϕ, *there exists a partial restriction* $\mathcal{M}''$ *of* $\mathcal{M}'$ *with respect to* $\neg\phi$ *such that* $(\mathcal{M}'', \emptyset, \prec_{AD}) \models INT_A(\phi)$.

[2] Note that a simple choice $buyPaperback \equiv buyMagazine$, on the other hand, will never be recovered unless he comes to believe it.

[3] In other words, as long as Agent A's belief monotonically expands by adopting a belief about ϕ and ϕ is consistent, he continues to intend ϕ, at least *conditionally*.

Furthermore, this property characterizes intentions among choices.

Theorem 13.4 $(\mathcal{M}, \emptyset, \prec_{AD}) \models INT_A(\phi)$ *iff*
$(\mathcal{M}, \emptyset, \prec_{AD}) \models CHO_A(\phi)$ *and for every partial restriction* $\mathcal{M}'$ *of* $\mathcal{M}$ *with respect to* ϕ*, there exists a partial restriction* $\mathcal{M}''$ *of* $\mathcal{M}'$ *with respect to* $\neg\phi$ *such that* $(\mathcal{M}'', \emptyset, \prec_{AD}) \models CHO_A(\phi)$.

13.4 Generalized Intention

In the preceding sections, we indentify intentions with desirable choices. But, we often use a word "intention" in a more general sense. For example, a natural language sentence "I want to go to a sushi bar." is said to express the speaker's intention of going to a sushi bar. However, going to a sushi bar is not necessarily desirable for him for itself. He may want go there only to achieve another goal, that is, eating sushi. In the terminology of planning theory, his going to a sushi bar is a subgoal of his eating sushi. Since this use of "intention" is also important for dialogue processing, we introduce a notion of generalized intention, which includes a notion of subgoal.

A subgoal of a goal is a sentence that must be satisfied in order to achieve that goal. For example, a precondition of an action is a subgoal of that action, and a conjunct of a conjunction is a subgoal of that conjunction. We formulate this notion as follows: Agent A thinks that ϕ is a subgoal of ψ (written $SBG_A(\phi, \psi)$) if he believes that ϕ is a necessary condition of ψ and ϕ is not temporally preceded by ψ. Since we consider attitudinal operators to be temporally neutral, the subgoal relation is defined only on sentences that contain no attitudinal operator. The precise definition is given as follows:

Definition 13.3 (Repeated) Let $S = (\mathcal{M}, \prec_{AP}, \prec_{AD})$ be an A-structure.

1. If ϕ or ψ contain attitudinal operators, $S \not\models SBG_A(\phi, \psi)$.
2. If ϕ contains no attitudinal operators and ψ is either p or $\neg p$, then $S \models SBG_A(\phi, \psi)$ iff
   (a) $S \models BEL_A(\psi \supset \phi)$, and
   (b) $time(q) \leq time(p)$ for all $q \in atom(\phi)$.
3. If ϕ, ψ and χ contain no attitudinal operators, then $S \models SBG_A(\phi, \psi \wedge \chi)$ iff there are ψ' and χ' such that
   (a) ϕ is logically equivalent to $\psi' \wedge \chi'$,
   (b) $atom(\phi) = atom(\psi') \cup atom(\chi')$, and
   (c) $M \models SBG_B(\psi', \psi) \wedge SBG_B(\chi', \chi)$.

4. If ϕ, ψ and χ contain no attitudinal operators, then
   $S \models SBG_A(\phi, \neg(\psi \wedge \chi))$ iff there are ψ' and χ' such that
   (a) ϕ is logically equivalent to $\neg(\psi' \wedge \chi')$,
   (b) $atom(\phi) = atom(\psi') \cup atom(\chi')$, and
   (c) $M \models SBG_B(\neg\psi', \neg\psi) \wedge SBG_B(\neg\chi', \neg\chi)$.

5. If ϕ contains no attitudinal operators, then
   $S \models SBG_A(\phi, \neg\neg\psi)$ iff $S \models SBG_A(\phi, \psi)$.

We can easily show the following properties of subgoal:

Theorem 13.5

1. $\models SBG_A(\phi, \psi) \supset BEL_A(\psi \supset \phi)$.
2. $\models SBG_A(\phi, \phi)$.
3. $\models SBG_A(\phi, \psi) \wedge SBG_A(\psi, \phi) \supset BEL_A(\phi \equiv \psi)$.
4. $\models SBG_A(\phi, \psi) \wedge SBG_A(\psi, \chi) \supset SBG_A(\phi, \chi)$.
5. $\models SBG_A(\phi, \psi) \wedge SBG_A(\phi', \psi') \supset SBG_A(\phi \wedge \phi', \psi \wedge \psi')$.
6. $\models SBG_A(\phi, \psi) \wedge SBG_A(\phi', \psi') \supset SBG_A(\phi \vee \phi', \psi \vee \psi')$.

A subgoal of a sentence is a necessary condition of that sentence. The subgoal relation is a partial order relation.

We say a sentence ϕ is a generalized intention if ϕ is not believed and ϕ is a subgoal of some intention:

Definition 13.4 (Repeated) $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models GINT_A(\phi)$ iff there is a ψ such that $(\mathcal{M}, \prec_{AP}, \prec_{AD}) \models \neg BEL_A(\phi) \wedge SBG_A(\phi, \psi) \wedge INT_A(\psi)$.

For example, when Agent A has an intention of eating sushi, he has generalized intention of going to a sushi bar:

$$\begin{aligned} \models\ & INT_A(eatSushi_2) \wedge \\ & BEL_A(eatSushi_2 \supset goSushiBar_1) \\ \supset\ & GINT_A(goSushiBar_1). \end{aligned}$$

We have the following properties of generalized intentions:

Theorem 13.6

1. *If ϕ contains no attitudinal operator,* $\models INT_A(\phi) \supset GINT_A(\phi)$.

2. $\models GINT_A(\phi) \supset CHO_A(\phi)$.

3. $\models GINT_A(\phi) \supset \neg BEL_A(\phi) \wedge \neg BEL_A(\neg\phi)$.

4. $\models GINT_A(\phi) \wedge GINT_A(\psi) \supset GINT_A(\phi \wedge \psi)$.

Generalized intentions are consistent and closed under conjunction. Furthermore, we can easily show that generalized intentions are also not closed under logical consequence.

Chapter 14

Reasoning about Plans

14.1 Basic Notions

In this chapter, we apply the preferential logic of mental attitudes to reasoning about plans. We analyze various examples of plan construction and plan recognition in our logic. Before we give examples, we need to introduce several basic notions about plans.

A notion of *plan* is widely used in AI in different ways. According to Pollack [32], there are two views of plans: plans as recipes for actions and plans as complex mental attitudes, and we deal with the latter of them. But, we will not further discuss what "plans" are. In fact, we regard plans simply as collections of an agent's intentions and generalized intentions, which are connected with each other by various relations such as subgoal, causation and equivalence. Some intentions are thought to be subgoals of other intentions. Some intentions are thought to cause other intentions. And, some intentions are believed to be equivalent to other intentions.

Action is a central notion for plans. Actions are specified in a way originated in the STRIPS system [13], using their *preconditions* and *effects*. Although notions of preconditions and effects of actions are ambiguous in both causal and temporal respects [32], we use these notions in later sections, and thus we define them informally here. We say a sentence ϕ is a precondition of another sentence (action) ψ if ψ implies ϕ and ϕ is temporally precedes ψ. We say ϕ is an effect of ψ if ψ implies ϕ and ψ is temporally precedes ϕ. For example, if an agent believes $eatSushi_2 \supset spendMoney_3 \wedge goSushiBar_1$, then he thinks that $goSushiBar_1$ is a precondition of $eatSushi_2$ and that $spendMoney_1$ is an effect of $eatSushi_2$.

Plan construction is the process of inferring intentions and generalized intentions in reverse temporal order, from top goals to subgoals and actions. An agent constructs his own plans and constructs other agents' plans (that is, simulates other agents' plan construction). *Plan recognition* is the process of inferring intentions and generalized intentions in temporal order, from observed actions and subgoals to top goals. Since we model plan construction and plan recognition in a single framework, we can use the same knowledge and preferences both in plan construction and in plan

recognition. For example, consider a preference of the form $D\text{-}PREF_A^{eq}(\neg act, act)$ where *act* is an occurrence of an action. It expresses that Agent A prefers not to perform *act* rather than to perform it if all else are equal, that is, if all effects of *act* are satisfied already. Since most actions are performed at some (physical or mental) cost, we think that an agent generally has this type of preference, and we can use it both in plan construction and in plan recognition.

14.2 Plan Construction

14.2.1 Introduction

To make the most desirable possible worlds true, an agent constructs a plan and coordinates his future actions. In this sense, an (ideal) agent always chooses a plan that is the most desirable for him. Since plans consist of intentions and generalized intentions, they are nonmonotonic. When an agent gets new pieces of information, his preferences for plans may change. Moreover, his plan may turn out to be impossible to perform. In those cases, the agent revises his plan to make it the most desirable for him.

There are a large number of plan construction models and planning systems [3, 7, 15]. Most of them have no means to express preference among plans, and thus they only concern with providing plans that can be adopted by an agent for given goals rather than with determining what plans are really adopted. There are several exceptions. The KAMP planning system [4] uses a *critic* procedure [35] that tries to find action subsumption to generate simpler (that is, more desirable) English sentences. The SUDO-PLANNER system [50] constructs plans under uncertainty. It chooses plausible plans by eliminating plans that are proved to be less likely to achieve given goals than other plans. However, only restricted types of preference can be represented in such systems. Furthermore, preferences are often represented implicitly in planning procedures, and it is hard to reason about them and their relationship with other attitudes.

Our logic provides a formal model of plan construction which has strong expressive power for preferences. An agent adopts desirable sentences as his goals (that is, intentions). Since properties are generally persistent (we express this fact by frame axioms), the agent needs to perform some action in order to achieve these goals. Then he chooses the most desirable actions that achieve them. In this way, an agent chooses his goals and plans for them which are the most desirable for him.

To illustrate this process, we give three examples of plan construction in the rest of this section. The first example illustrates plan construction and revision with conflicting preferences. In the second example, we deal with simulation of another agent's plan construction processes. In the third example, we deal with cooperative plan construction, that is, construction of plans that achieve another agent's goals. In these examples, we specify an agent's mental states by structure specification lists. To ensure that preference orders are bounded, we assume that $atom(\mathcal{L})$ is

finite in this section.

14.2.2 Constructing a Plan with Multiple Preferences

Consider that Agent A is hungry and plans to eat lunch. He knows two restaurants, a sushi bar where he can eat sushi, ptand a noodle shop where he can eat soba and udon. Agent A has several preferences which conflict with each other. First of all, he wants to be not hungry. This preference has the highest priority. He prefers to eat sushi rather than to eat soba, and he prefers soba to udon. But, he does not want to eat lunch at a crowded restaurant. This situation is described by the following structure specification list:

$$
\begin{aligned}
\mathcal{K} = \{ & (1)\, goSushiBar_1 \supset atSushiBar_2, \\
& (2)\, goNoodleShop_1 \supset atNoodleShop_2, \\
& (3)\, eat_3 \equiv eatSushi_3 \vee eatSoba_3 \vee eatUdon_3, \\
& (4)\, eatSushi_3 \supset atSushiBar_2, \\
& (5)\, eatSoba_3 \vee eatUdon_3 \supset atNoodleShop_2, \\
& (6)\, eat_3 \supset \neg hungry_4, \\
& (7)\, atCrowded_2 \equiv (atSushiBar_2 \wedge crowdedSushiBar_2) \vee \\
& \qquad\qquad (atNoodleShop_2 \wedge crowdedNoodleShop_2), \\
& (8)\, \neg atSushiBar_0, \\
& (9)\, \neg atNoodleShop_0, \\
& (10)\, hungry_0 \\
\} & \\
\mathcal{P} = \langle & (11)\, P\text{-}PREF_A^{0,2}(\neg atSushiBar_2, atSushiBar_2 \mid \neg atSushiBar_0), \\
& (12)\, P\text{-}PREF_A^{0,2}(\neg atNoodleShop_2, atNoodleShop_2 \mid \neg atNoodleShop_0), \\
& (13)\, P\text{-}PREF_A^{0,4}(hungry_4, \neg hungry_4 \mid hungry_0) \\
\rangle & \\
\mathcal{D} = \langle & (14)\, D\text{-}PREF_A^{\emptyset}(\neg hungry_4, hungry_4), \\
& (15)\, D\text{-}PREF_A^{\emptyset}(\neg atCrowded_2, atCrowded_2), \\
& (16)\, D\text{-}PREF_A^{\emptyset}(eatSushi_3, eatSoba_3), \\
& (17)\, D\text{-}PREF_A^{\emptyset}(eatSoba_3, eatUdon_3) \\
\rangle &
\end{aligned}
$$

$\mathcal{K}$ consists of Agent A's knowledge about general properties of actions and facts about the current situation: he is hungry and not at restaurant at time 0. $\mathcal{P}$ consists of frame axioms. For example, Sentence 11 expresses that if Agent A is not at a sushi bar at time 0, then his being not at the sushi bar at time 2 is more plausible than his being at the sushi bar if states of the world between time 0 and time 2 are equal. We think that Agent A has such frame axioms for every interval of time, though here we give only those which we really use. $\mathcal{D}$ specifies Agent A's preferences about desirability. Sentence 14 has the highest priority, Sentence 15 is the second, and Sentence 16 is the third. Then, we have the following result:

Theorem 14.1

$$\begin{aligned}(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models\ & BEL_A(\neg hungry_4 \supset eat_3) \wedge \\ & INT_A(\neg hungry_4 \wedge \neg atCrowded_2) \wedge \\ & INT_A(\neg hungry_4) \wedge \\ & INT_A(eat_3) \wedge \\ & INT_A(eatSushi_3) \wedge \\ & GINT_A(atSushiBar_2) \wedge \\ & GINT_A(goSushiBar_1) \wedge \\ & GINT_A(\neg atCrowded_2).\end{aligned}$$

Agent A intends to be not hungry and not to eat at a crowded restaurant. Since he does not believe that these restaurants are crowded, he chooses to eat sushi and plans to go to the sushi bar.

Before leaving for the sushi bar, Agent A happens to know that the bar is crowded:

$$\mathcal{K}' = \mathcal{K} \cup \{crowdedSushiBar_2\}.$$

His intention of eating sushi becomes inconsistent with a stronger intention not to eat at a crowded restaurant, and thus it is abandoned. In other words, he prefers to eat at the noodle shop rather than to eat at the crowded sushi bar and revises his plan as follows:

Theorem 14.2

$$\begin{aligned}(Mod(\mathcal{K}'), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models\ & INT_A(eatSoba_3) \wedge \\ & GINT_A(atNoodleShop_2) \wedge \\ & GINT_A(goNoodleShop_1) \wedge \\ & GINT_A(\neg atSushiBar_2) \wedge \\ & GINT_A(\neg goSushiBar_1).\end{aligned}$$

At the noodle shop, Agent A is told that soba is out of stock:

$$\mathcal{K}'' = \mathcal{K}' \cup \{\neg eatSoba_3\}.$$

His plan is now impossible to perform, and he constructs a new plan to eat udon:

Theorem 14.3

$$(Mod(\mathcal{K}''), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models INT_A(eatUdon_3).$$

14.2.3 Simulating Plan Construction of Another Agent

An agent constructs not only his own plans but also other agents' plans using knowledge about their belief and preferences. If he has complete knowledge about them, he gets as strong conclusions as for his own plans. If he has only incomplete knowledge, he gets weaker conclusions. The following example demonstrates how our logic models this capability of simulating other agents' plan construction.

We use the same lunch situation except that the actor is Agent B. We transform the previous structure specification list into a new one in the following way: First, to make the new actor clear, we add "B" to the names of atoms and attitudinal operators. Next, we apply an operator BEL_B to every sentence in $\mathcal{K}$. For frame axioms in $\mathcal{P}$, we replace them for brevity by beliefs they actually express in this situation and add them to $\mathcal{K}$. We transform $\mathcal{D}$ into an equivalent compatible set $C(\mathcal{D})$ by using the method explained in Section 12.5. Then, we have the following list:

$$
\begin{aligned}
\mathcal{K} = \{ & (1)\ BEL_B(goSushiBarB_1 \supset atSushiBarB_2), \\
& (2)\ BEL_B(goNoodleShopB_1 \supset atNoodleShopB_2), \\
& (3)\ BEL_B(eatB_3 \equiv eatSushiB_3 \vee eatSobaB_3 \vee eatUdonB_3), \\
& (4)\ BEL_B(eatSushiB_3 \supset atSushiBarB_2), \\
& (5)\ BEL_B(eatSobaB_3 \vee eatUdonB_3 \supset atNoodleShopB_2), \\
& (6)\ BEL_B(eatB_3 \supset \neg hungryB_4), \\
& (7)\ BEL_B(atCrowdedB_2 \equiv (atSushiBarB_2 \wedge crowdedSushiBar_2) \vee \\
& \qquad\qquad (atNoodleShopB_2 \wedge crowdedNoodleShop_2)), \\
& (8)\ BEL_B(\neg atSushiBarB_0), \\
& (9)\ BEL_B(\neg atNoodleShopB_0), \\
& (10)\ BEL_B(hungryB_0), \\
& (11)\ BEL_B(\neg atSushiBarB_0 \wedge atSushiBarB_2 \supset goSushiBarB_1), \\
& (12)\ BEL_B(\neg atNoodleShopB_0 \wedge atNoodleShopB_2 \supset goNoodleShopB_1), \\
& (13)\ BEL_B(hungryB_0 \wedge \neg hungryB_4 \supset eatB_3), \\
& (14)\ D\text{-}PREF_B^{\emptyset}(\neg hungryB_4, hungryB_4), \\
& (15)\ D\text{-}PREF_B^{\emptyset}(\neg atCrowdedB_2, atCrowdedB_2 \mid \neg hungryB_4), \\
& (16)\ D\text{-}PREF_B^{\emptyset}(\neg atCrowdedB_2, atCrowdedB_2 \mid hungryB_4), \\
& (17)\ D\text{-}PREF_B^{\emptyset}(eatSushiB_3, eatSobaB_3 \mid \neg hungryB_4 \wedge \neg atCrowdedB_2), \\
& (18)\ D\text{-}PREF_B^{\emptyset}(eatSushiB_3, eatSobaB_3 \mid \neg hungryB_4 \wedge atCrowdedB_2), \\
& (19)\ D\text{-}PREF_B^{\emptyset}(eatSushiB_3, eatSobaB_3 \mid hungryB_4 \wedge \neg atCrowdedB_2), \\
& (20)\ D\text{-}PREF_B^{\emptyset}(eatSushiB_3, eatSobaB_3 \mid hungryB_4 \wedge atCrowdedB_2), \\
& (21)\ D\text{-}PREF_B^{\emptyset}(eatSobaB_3, eatUdonB_3 \mid \neg hungryB_4 \wedge \neg atCrowdedB_2 \wedge \\
& \qquad eatSushiB_3), \\
& (22)\ D\text{-}PREF_B^{\emptyset}(eatSobaB_3, eatUdonB_3 \mid \neg hungryB_4 \wedge \neg atCrowdedB_2 \wedge \\
& \qquad \neg eatSushiB_3), \\
& (23)\ D\text{-}PREF_B^{\emptyset}(eatSobaB_3, eatUdonB_3 \mid \neg hungryB_4 \wedge atCrowdedB_2 \wedge \\
& \qquad eatSushiB_3),
\end{aligned}
$$

(24) $D\text{-}PREF^{\emptyset}_{B}(eatSobaB_3, eatUdonB_3 \mid \neg hungryB_4 \wedge atCrowdedB_2 \wedge \neg eatSushiB_3)$,
(25) $D\text{-}PREF^{\emptyset}_{B}(eatSobaB_3, eatUdonB_3 \mid hungryB_4 \wedge \neg atCrowdedB_2 \wedge eatSushiB_3)$,
(26) $D\text{-}PREF^{\emptyset}_{B}(eatSobaB_3, eatUdonB_3 \mid hungryB_4 \wedge \neg atCrowdedB_2 \wedge \neg eatSushiB_3)$,
(27) $D\text{-}PREF^{\emptyset}_{B}(eatSobaB_3, eatUdonB_3 \mid hungryB_4 \wedge atCrowdedB_2 \wedge eatSushiB_3)$,
(28) $D\text{-}PREF^{\emptyset}_{B}(eatSobaB_3, eatUdonB_3 \mid hungryB_4 \wedge atCrowdedB_2 \wedge \neg eatSushiB_3)$

}

$\mathcal{P} = \langle\,\rangle$

$\mathcal{D} = \langle\,\rangle$

We let $\mathcal{D}$ be an empty set, because Agent A's desire is not the matter here. We have the following:

Theorem 14.4

$$
\begin{aligned}
(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models\ & BEL_A((C1 \supset INT_B(\neg hungryB_4)) \wedge \\
& (C1 \supset INT_B(eatB_3)) \wedge \\
& (C1 \wedge C2 \supset INT_B(eatSushiB_3)) \wedge \\
& (C1 \wedge C2 \wedge C3 \supset GINT_B(atSushiBarB_2)) \wedge \\
& (C1 \wedge C2 \wedge C3 \supset GINT_B(goSushiBarB_1)) \wedge \\
& (C1 \wedge C4 \supset INT_B(eatSobaB_3)) \wedge \\
& (C1 \wedge C4 \wedge C5 \supset GINT_B(atNoodleShopB_2)) \wedge \\
& (C1 \wedge C4 \wedge C5 \supset GINT_B(goNoodleShopB_1)))
\end{aligned}
$$

where

$$
\begin{aligned}
C1 &= \neg BEL_B(hungryB_4) \wedge \neg BEL_B(\neg hungryB_4), \\
C2 &= \neg BEL_B(eatSushiB_3 \wedge \neg eatSobaB_3 \supset atCrowdedB_2) \wedge \\
&\quad \neg BEL_B(\neg eatSushiB_3 \wedge eatSobaB_3 \wedge \neg eatUdonB_3 \supset atCrowdedB_2), \\
C3 &= \neg BEL_B(atSushiBarB_2), \\
C4 &= BEL_B(eatSushiB_3 \wedge \neg eatSobaB_3 \supset atCrowdedB_2) \wedge \\
&\quad \neg BEL_B(\neg eatSushiB_3 \wedge eatSobaB_3 \wedge \neg eatUdonB_3 \supset atCrowdedB_2), \\
C5 &= \neg BEL_B(atNoodleShopB_2).
\end{aligned}
$$

We have an weak result with conditions on Agent B's belief. It is because we use only positive knowledge about Agent B, that is, knowledge about what B believes and prefers, and we does not use knowledge about what B does not believe and does not prefer. In general, an agent thinks by default that another agent does not believe ϕ, for every ϕ that belongs to a certain class of sentences. For simplicity, we assume this type of preference for all propositional sentences α:

$$\mathcal{P}' = \{\, P\text{-}PREF_A^{\emptyset}(\neg BEL_B(\alpha), BEL_B(\alpha)) \mid \alpha \text{ is a propositional sentence}\}$$

where the precedence order on $\mathcal{P}'$ is an arbitrary well-order. Then, we have the following:

Theorem 14.5

$$\begin{aligned}(Mod(\mathcal{K}), \prec_{\mathcal{P}'}, \prec_{\mathcal{D}}) \models\ & BEL_A(INT_B(\neg hungryB_4)) \wedge \\ & INT_B(eatB_3) \wedge \\ & INT_B(eatSushiB_3) \wedge \\ & GINT_B(atSushiBarB_2) \wedge \\ & GINT_B(goSushiBarB_1)).\end{aligned}$$

14.2.4 Cooperative Plan Construction

An agent often constructs plans that achieve another agent's goals. When an agent believes that another agent has an intention, he usually intends to achieve that intention for that agent. In Chapter 6 of Part 1, we express this fact by an inheritance rule (C9). This rule says, in terms of our logic, that we can infer $INT_A(\alpha)$ from $BEL_A(INT_B(\alpha))$. However, this inference is a default one, and it is not always applicable. In particular, this inference is not valid when Agent A believes either α or $\neg\alpha$ and when Agent A has stronger preferences that conflict with α.

In our logic, we can deal with this type of inference using preferences of the following form:

$$D\text{-}PREF_A^{\emptyset}(\alpha, \neg\alpha \mid INT_B(\alpha)).$$

Consider that Agent A knows that his companion Agent B wants to go to a museum. To go to the museum, B needs to know where the bus for the museum starts, and A knows it. This situation is described by the following structure specification list:

$$\begin{aligned}\mathcal{K} = \{\ & (1)\ informGateA_1 \supset knowGateB_2, \\ & (2)\ gotoMuseumB_3 \supset knowGateB_2 \wedge atMuseumB_4, \\ & (3)\ enterMuseumB_5 \supset atMuseumB_4 \wedge openMuseum_4, \\ & (4)\ \neg knowGateB_0, \\ & (5)\ \neg atMuseumB_0, \\ & (6)\ INT_B(enterMuseumB_5) \\ \}\ & \\ \mathcal{P} = \langle\ & (7)\ P\text{-}PREF_A^{0,2}(\neg knowGateB_2, knowGateB_2 \mid \neg knowGateB_0), \\ & (8)\ P\text{-}PREF_A^{0,4}(\neg at, atMuseumB_4 \mid \neg atMuseumB_0) \\ \rangle\ & \\ \mathcal{D} = \langle\ & (9)\ D\text{-}PREF_A^{\emptyset}(watchParadeB_6, \neg watchParadeB_6), \\ & (10)\ D\text{-}PREF_A^{\emptyset}(enterMuseumB_5, \neg enterMuseumB_5 \mid INT_B(enterMuseumB_5)) \\ \rangle\ & \end{aligned}$$

In this situation, Agent A inherits B's intention of entering the museum and constructs a plan for it. He intends to inform B of where the bus starts:

Theorem 14.6

$$\begin{aligned}(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models\ & BEL_A(knowGateB_2 \supset informGateA_1) \wedge \\ & BEL_A(atMuseumB_4 \supset gotoMuseumB_3) \wedge \\ & INT_A(watchParadeB_6) \wedge \\ & INT_A(enterMuseumB_5) \wedge \\ & GINT_A(atMuseumB_4) \wedge \\ & GINT_A(gotoMuseumB_3) \wedge \\ & GINT_A(knowGateB_2) \wedge \\ & GINT_A(informGateA_1) \wedge \\ & GINT_A(openMuseum_4).\end{aligned}$$

On the other hand, if Agent A knows that the museum is not open

$$\mathcal{K}' = \mathcal{K} \cup \{\neg openMuseum_4\}$$

and B's intention of entering it is not achievable, he does not inherit this intention. Similarly, if this intention conflicts with A's stronger intention, say, an intention of letting B watch the city parade

$$\mathcal{K}'' = \mathcal{K} \cup \{enterMuseumB_5 \supset \neg watchParadeB_6\}.$$

it does not inherited.

Theorem 14.7

$$\begin{aligned}(Mod(\mathcal{K}'), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) &\models \neg INT_A(enterMuseumB_5). \\ (Mod(\mathcal{K}''), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) &\models \neg INT_A(enterMuseumB_5).\end{aligned}$$

14.3 Plan Recognition

14.3.1 Introduction

In multi-agent environments, an agent recognizes other agents' intentions and plans in order to coordinate his actions with them and to behave cooperatively. Plan recognition is particularly important when we want to construct a dialogue system that generates helpful responses, and it is intensively studied in relation to dialogue understanding.

Procedural models of plan recognition [1, 6, 24] infer plausible plans of another agent from observed actions or intentions of that agent by using heuristic rules and

inference procedures. Two heuristic rules are widely used [1]: the *Action-Effect Rule* says that if an agent intends to perform an action, it is plausible that he intends to achieve its effects. The *Precondition-Action Rule* says that if an agent intends to achieve a precondition of an action, it is plausible that he intends to perform that action.

Formal studies of plan recognition, which give theoretical foundations to these procedural models, have not made much so far. Kautz [23] applied *circumscription* [28] to plan recognition in domains where actions are hierarchically structured with respect to decomposition and abstraction. He gave a nice account of plan recognition in these restricted domains. However, it is hard to extend this model to allow various domain-specific preferences as inputs, since circumscription deals with only restricted form of preferences.

In this section, we apply our logic of mental attitudes to modeling plan recognition. We can specify arbitrary preferences between sentences as inputs to the recognition. In fact, we can use two kinds of preferences: a planner's preference about desirability, and a recognizer's preferences about the planner's mental attitudes.

To examine how such preferences can be used to explain typical phenomena in plan recognition, we give four examples in the rest of this section. The first and the second examples examine the Action-Effect Rule and the Precondition-Action Rule, respectively. We show that under certain conditions on preferences, these rules are valid. In the third example, we deal with an extension of the Action-Effect Rule. When an intended action has several effects, the most preferred effects are considered to be intended. The fourth example deals with a recognizer's preference for simpler plans. We give comparatively small examples here. In order to explain larger examples, we need many preferences which are not necessarily easy to understand. Therefore, we need to know much about regularities in an agent's preferences. In examples presented in this section, we assume that $atom(\mathcal{L})$ includes no atom other than those occurring in the structure specification lists.

14.3.2 Inferring Effects from Actions

Consider that Agent A knows that Agent B intends to perform an action, an action of his going to a museum by bus. A knows that B can get to the museum by performing this action or another action, going there on foot. A knows that B does not prefer to go to the museum if other things are equal (that is, B does not prefer to perform a going action if he is already at the museum). This situation is described as follows:

$$
\begin{aligned}
\mathcal{K} = \{ & (1)\ BEL_B(gobyBusB_2 \vee goOnFootB_2 \supset atMuseumB_3), \\
& (2)\ D\text{-}PREF_B^{eq}(\neg gobyBusB_2, gobyBusB_2), \\
& (3)\ D\text{-}PREF_B^{eq}(\neg goOnFootB_2, goOnFootB_2), \\
& (4)\ D\text{-}PREF_B^{eq}(\neg gobyBusB_2 \wedge \neg goOnFootB_2, gobyBusB_2 \wedge goOnFootB_2), \\
& (5)\ INT_B(gobyBusB_2) \\
\} & \\
\mathcal{P} = \langle\, \rangle & \\
\mathcal{D} = \langle\, \rangle &
\end{aligned}
$$

Then we can infer B's intention of achieving the effect of the going action, that is, his intention of being at the museum:

Theorem 14.8

$$
\begin{aligned}
& (Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models \\
& \quad BEL_A(\ BEL_B(atMuseumB_3 \supset gobyBusB_2 \vee goOnFootB_2) \wedge \\
& \qquad (\neg BEL_B(gobyBusB_2 \supset \neg goOnFootB_2) \supset INT_B(atMuseumB_3)) \wedge \\
& \qquad D\text{-}PREF_B^{eq}(gobyBusB_2, goOnFootB_2)).
\end{aligned}
$$

In fact, we get information about Agent B's preference, that he prefers to go by bus rather than to go on foot. This example shows that agents' preferences not only serve as inputs to plan recognition, but also they are sometimes obtained as a part of the output of it.

14.3.3 Inferring Actions from Preconditions

Consider that Agent B has a generalized intention of knowing where the bus starts, which is a precondition of his going somewhere by bus.

$$
\begin{aligned}
\mathcal{K} = \{ & (1)\ BEL_B(gobyBusB_2 \supset knowGateB_1), \\
& (2)\ GINT_B(knowGateB_1) \\
\} & \\
\mathcal{P} = \langle & (3)\ P\text{-}PREF_A^{\emptyset}(\neg INT_B(knowGateB_1), INT_B(knowGateB_1)), \\
& (4)\ P\text{-}PREF_A^{\emptyset}(INT_B(gobyBusB_2), INT_B(knowGateB_1 \wedge \neg gobyBusB_2)) \\
\rangle & \\
\mathcal{D} = \langle\, \rangle &
\end{aligned}
$$

Note that knowing where the bus starts is usually not desirable for itself, and thus it is not an intention as is expressed by Sentence 3. Sentence 4 says that a simple intention $INT_B(gobyBusB_2)$ is more likely to be adopted than a complex and unnatural intention $INT_B(knowGateB_1 \wedge \neg gobyBusB_2)$. Then we can conclude that Agent B has the generalized intention $knowGateB_1$ simply as a subgoal of his intention of performing an action $gobyBusB_2$:

Theorem 14.9

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models BEL_A(INT_B(gobyBusB_2) \wedge \neg INT_B(knowGateB_1)).$$

To get the similar result, we can use another set of preferences:

$$\mathcal{P}' = \langle P\text{-}PREF_A^{\emptyset}(INT_B(gobyBusB_2), \neg INT_B(gobyBusB_2) \mid GINT_B(knowGateB_1)) \rangle$$

This preference directly expresses our heuristics for plan recognition:
from $GINT_B(knowGateB_1)$ infer $INT_B(gobyBusB_2)$.

Theorem 14.10

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}'}, \prec_{\mathcal{D}}) \models BEL_A(INT_B(gobyBusB_2)).$$

14.3.4 Actions with Several Effects

We have seen that we can infer an intention of achieving effects of an action from an intention of performing that action using a certain set of preferences. In situations where an intended action has several effects, we need to choose which effect is intended. For example, an action of going to a museum by bus has two effects, his being at the museum and his spending money on the bill. Since the latter effect is clearly not desirable, we can easily choose the intended effect.

$$\begin{aligned}
\mathcal{K} = \{ & (1)\ BEL_B(gobyBusB_2 \supset atMuseumB_3 \wedge spendMoneyB_3), \\
& (2)\ D\text{-}PREF_B^{eq}(\neg gobyBusB_2, gobyBusB_2), \\
& (3)\ D\text{-}PREF_B^{eq}(\neg gobyBusB_2 \wedge \neg spendMoneyB_3, gobyBusB_2), \\
& (4)\ INT_B(gobyBusB_2) \\
\} & \\
\mathcal{P} = \langle\, \rangle & \\
\mathcal{D} = \langle\, \rangle &
\end{aligned}$$

We have the following:

Theorem 14.11

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models BEL_A(INT_B(atMuseumB_3 \wedge spendMoneyB_3) \wedge INT_B(atMuseumB_3)).$$

In many cases, we do not have B's desirability preferences like above. Nevertheless, we can choose among effects if we have plausibility preferences about B's intentions:

$$
\begin{aligned}
\mathcal{K}' = \{ & (1)\ BEL_B(gobyBusB_2 \supset atMuseumB_3 \wedge spendMoneyB_3), \\
& (2)\ D\text{-}PREF_B^{eq}(\neg gobyBusB_2, gobyBusB_2), \\
& (4)\ INT_B(gobyBusB_2) \\
\}& \\
\mathcal{P}' = \langle & (5)\ P\text{-}PREF_A^{\emptyset}(INT_B(atMuseumB_3) \vee INT_B(spendMoneyB_3), \\
& \qquad \neg(INT_B(atMuseumB_3) \vee INT_B(spendMoneyB_3)) \mid \\
& \qquad INT_B(atMuseumB_3 \wedge spendMoneyB_3)), \\
& (6)\ P\text{-}PREF_A^{\emptyset}(INT_B(atMuseumB_3), INT_B(spendMoneyB_3)) \\
\rangle & \\
\mathcal{D}' = \langle & \rangle
\end{aligned}
$$

Then we have the following:

Theorem 14.12

$$
(Mod(\mathcal{K}'), \prec_{\mathcal{P}'}, \prec_{\mathcal{D}'}) \models BEL_A(INT_B(atMuseumB_3 \wedge spendMoneyB_3) \wedge INT_B(atMuseumB_3)).
$$

14.3.5 Preference for Simpler Plans

Consider a situation taken from [23]: Agent A knows that Agent B has two intention, an intention of getting a gun and an intention of going to a bank. B intends to get a gun only when he intends to go hunting or he intends to rob a bank. He intends to go to the bank only when he intends to check cash or he intends to rob the bank.

$$
\begin{aligned}
\mathcal{K} = \{ & (1)\ INT_B(getGunB_1) \supset INT_B(huntB_3) \vee INT_B(robBankB_3), \\
& (2)\ INT_B(gotoBankB_2) \supset INT_B(cashCehckB_3) \vee INT_B(robBankB_3), \\
& (3)\ INT_B(getGunB_1), \\
& (4)\ INT_B(gotoBankB_2) \\
\}& \\
\mathcal{P} = \langle & \rangle \\
\mathcal{D} = \langle & \rangle
\end{aligned}
$$

Then we have the following:

Theorem 14.13

$$
(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models BEL_A((INT_B(huntB_3) \wedge INT_B(cashCheckB_3)) \vee INT_B(robBankB_3)).
$$

There are two possible plans for B, one consists of an intention of hunting and an intention of checking cash, and the other consists of a single intention of robbing the bank. Usually, the latter plan is preferred, since it contains fewer actions. A preference for simpler plans is called *Occam's razor* and used widely. It is particularly important in dialogue understanding where coherence of utterances is assumed.

Kautz [23] formulated this preference by extending circumscription. He minimizes the number of *End* actions in plans, where an End action is an action that is not a component of any other actions. In our logic, this preference is expressed by sentences of the following form:

$$P\text{-}PREF_A^{\emptyset}(INT_B(E1) \wedge \cdots \wedge INT_B(En), INT_B(E'1) \wedge \cdots \wedge INT_B(E'm))$$

where $E1, \cdots, En, E'1, \cdots, E'm$ are distinct End actions and $n < m$. Our example has three End actions, $huntB_3$, $cashCheckB_3$ and $robBankB_3$, and thus we get the following list:

$$\begin{aligned}\mathcal{P}' = \langle\ & P\text{-}PREF_A^{\emptyset}(INT_B(robBankB_3), INT_B(huntB_3) \wedge INT_B(cashCheckB_3)),\\ & P\text{-}PREF_A^{\emptyset}(INT_B(huntB_3), INT_B(cashCheckB_3) \wedge INT_B(robBankB_3)),\\ & P\text{-}PREF_A^{\emptyset}(INT_B(cashCheckB_3), INT_B(robBankB_3) \wedge INT_B(huntB_3))\\ \rangle\end{aligned}$$

The precedence order is not the matter here. Now we can choose the simpler plan:

Theorem 14.14

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}'}, \prec_{\mathcal{D}}) \models BEL_A(INT_B(robBankB_3)).$$

This general preference for simpler plans can be overridden by specific domain-dependent preferences. For example, consider that Agent A knows that Agent B likes hunting, and he thinks that it is more plausible for B to intend hunting than to intend robbing a bank. We assume this preference has higher priority than his preference for simpler plans:

$$\begin{aligned}\mathcal{P}'' = \langle\ & P\text{-}PREF_A^{\emptyset}(INT_B(huntB_3), INT_B(robBankB_3)),\\ & P\text{-}PREF_A^{\emptyset}(INT_B(robBankB_3), INT_B(huntB_3) \wedge INT_B(cashCheckB_3)),\\ & P\text{-}PREF_A^{\emptyset}(INT_B(huntB_3), INT_B(cashCheckB_3) \wedge INT_B(robBankB_3)),\\ & P\text{-}PREF_A^{\emptyset}(INT_B(cashCheckB_3), INT_B(robBankB_3) \wedge INT_B(huntB_3))\\ \rangle\end{aligned}$$

Then, we get a different conclusion as follows:

Theorem 14.15

$$\begin{aligned}(Mod(\mathcal{K}), \prec_{\mathcal{P}''}, \prec_{\mathcal{D}}) \models\ & BEL_A(INT_B(huntB_3) \wedge\\ & (INT_B(cashCheckB_3) \vee INT_B(robBankB_3)).\end{aligned}$$

Chapter 15

Conclusion

We have presented a model of an agent in the form of a logic of mental attitudes based on preference ordering. We have dealt with qualitative preferences, which are explicitly represented by partial orders on model structures. An agent's mental state is specified by knowledge and two preference orders, that is, the plausibility order and the desirability order. The language of our logic is an extended propositional language with attitudinal operators: belief, intention, choice, and preference between sentences. We define the satisfaction relation for these operators in terms of the preference orders. Furthermore, we have introduced a construct of sentences, which is used to specify an agent's knowledge and the preference orders. We have applied this logic to reasoning about plans. We have given a formal account of plan construction and selection process, and examined several heuristics for plan recognition currently used.

The main contributions of Part 2 are as follows:

1. We have presented a model of an agent based on two preference orders: one is about plausibility and the other is about desirability. Mental attitudes such as belief, intention and choice have been defined in terms of preference orders.

2. We have introduced operators for preference between sentences, and examined their properties. We have given several types of preferences frequently used, and have shown that the strongest type of them is closely related to other mental attitudes such as belief and intention.

3. We have given an intuitive formulation of a notion of intention which satisfies most of requisites ever proposed such as freedom from consequential closure and persistency.

4. We have given a formal account of plan construction and selection processes. We have modeled plan construction with multiple preferences and dynamic revision of plans.

5. We have examined several heuristics for plan recognition currently used. We have given preferences that validate widely used heuristics, and demonstrated that our framework can give a good formal basis for plan recognition models.

Appendix A

Proofs of Theorems

Theorem 12.1

1. *P-PREF$_A^T$ is monotonic, and*
2. *the other operators BEL$_A$, CHO$_A$, D-PREF$_A^T$, INT$_A$, SBG$_A$ and GINT$_A$ are nonmonotonic.*

Proof:

1. Directly follows from the definition.
2. We give a counterexample as follows: Let p and q be distinct atoms of $\mathcal{L}$ that satisfy $time(p) = time(q)$. We take

$$\begin{aligned} \mathcal{M}_1 &= \{M \mid M \models p \wedge \neg q\} \\ \mathcal{M}_2 &= \{M \mid M \models p \wedge q\} \\ \mathcal{M}_3 &= \{M \mid M \models \neg p\} \end{aligned}$$

and define orders $\prec_{AP}$, $\prec_{AD}$by

$$\begin{aligned} M_1 \prec_{AP} M_2 \quad &\text{iff} \quad M_1 \in \mathcal{M}_3 \text{ and} M_2 \in \mathcal{M}_1 \cup \mathcal{M}_2. \\ M_1 \prec_{AD} M_2 \quad &\text{iff} \quad M_1 \in \mathcal{M}_1 \text{ and} M_2 \in \mathcal{M}_2. \end{aligned}$$

Then, $(\mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3, \prec_{AP}, \prec_{AD})$ satisfies $BEL_A(p)$, $CHO_A(q)$, $D\text{-}PREF_A^T(q, \neg q)$, $INT_A(q)$, $SBG_A(p \wedge q, q)$ and $GINT_A(p \wedge q)$, but $(\mathcal{M}_3, \prec_{AP}, \prec_{AD})$ satisfies none of them.

□

Theorem 12.3 *For a belief set $K = B_{\mathcal{M}}$ and sentences ϕ and ψ such that $\mathcal{M} \cap Mod(\phi \wedge \psi)$ is nonempty, the revision function $*$ satisfies Gärdenfors's postulates (K*1),(K*2),(K*3),(K*5),(K*6) and (K*7).*

Proof: (K^*1) follows from the fact that beliefs are closed under logical consequence in our logic. (K^*2) is true because $(\mathcal{M} \cap Mod(\phi), \prec_{AP}, \prec_{AD})$ obviously satisfies $BEL_A(\phi)$. To prove (K^*3), it is sufficient to show that if $M \in Max(\mathcal{M}, \prec_{AP})$ and $M \models \phi$ then $M \in Max(\mathcal{M} \cap Mod(\phi), \prec_{AP})$, which can be easily proved. (K^*5) holds trivially, since both sides of the condition are always true. (K^*6) is true because $Mod(\phi) = Mod(\psi)$. (K^*7) is a special case of (K^*3) and thus satisfied, since $K^*_{\phi \wedge \psi} = (K^*_\phi)^*_\psi$ by our definition. □

Theorem 12.9

1. *$\prec_{\mathcal{X}}$ is a strict partial order.*
2. *$\prec_{\mathcal{X}} \subset (\bigcup_{X \in \mathcal{X}} \prec X)^+$.*
3. *If $\mathcal{X}$ is compatible, $\prec_{\mathcal{X}} = (\bigcup_{X \in \mathcal{X}} \prec X)^+$.*
4. *If $\mathcal{X}'$ is an initial segment of $\mathcal{X}$, $\prec_{\mathcal{X}'} \subset \prec_{\mathcal{X}}$.*

Proof:

1. We only need to show that $\prec_{\mathcal{X}}$ is irreflexive. Assume that $\prec_{\mathcal{X}}$ is not irreflexive. Then there exist B-structures $M_1, \ldots, M_n$ and sentences $X_1, \ldots, X_n \in \mathcal{X}$ such that $M_1 \overline{\prec X_1} M_2 \overline{\prec X_2} \cdots \overline{\prec X_{n-1}} M_n \overline{\prec X_n} M_1$. Without loss of generality, we may assume that X_n is preceded by all of $X_1, \ldots$, and X_{n-1}. Then $(\prec X_n \cup \bigcup_{X precedes X_n} \overline{\prec X})^+$ includes (M_1, M_n), and thus (M_n, M_1) is removed from $\overline{\prec X_n}$, which contradicts our assumption.
2. Since $\overline{\prec X} \subset \prec X$ for all $X \in \mathcal{X}$, we have $\prec_{\mathcal{X}} = (\bigcup_{X \in \mathcal{X}} \overline{\prec X})^+ \subset (\bigcup_{X \in \mathcal{X}} \prec X)^+$.
3. We prove $\overline{\prec X} = \prec X$ for all $X \in \mathcal{X}$ by transfinite induction. Assume that $\overline{\prec X'} = \prec X'$ for all $X' \in \mathcal{X}$ that precedes X. Since it is clear that a subset of a compatible set is also compatible, $(\prec X \cup \bigcup_{X' precedes X} \prec X')^+$ is irreflexive, and thus $((\prec X \cup \bigcup_{X' precedes X} \prec X')^+)^{-1}$ and $\prec X$ have no common element. It follows that $\overline{\prec X} = \prec X$.
4. Obvious from the definition of $\prec_{\mathcal{X}}$.

□

Theorem 13.3 *If $(\mathcal{M}, \emptyset, \prec_{AD}) \models INT_A(\phi)$, then for every partial restriction $\mathcal{M}'$ of $\mathcal{M}$ with respect to ϕ, there exists a partial restriction $\mathcal{M}''$ of $\mathcal{M}'$ with respect to $\neg\phi$ such that $(\mathcal{M}'', \emptyset, \prec_{AD}) \models INT_A(\phi)$.*

Proof: Let $\mathcal{M}'' = \{M \in \mathcal{M}' \mid M \in Mod(\phi)$ or there exists $M' \in \mathcal{M}' \cap Mod(\phi)$ such that $M \prec_{AD} M'\}$. □

Theorem 13.4 $(\mathcal{M}, \emptyset, \prec_{AD}) \models INT_A(\phi)$ *iff*
$(\mathcal{M}, \emptyset, \prec_{AD}) \models CHO_A(\phi)$ *and for every partial restriction* $\mathcal{M}'$ *of* $\mathcal{M}$ *with respect to* ϕ*, there exists a partial restriction* $\mathcal{M}''$ *of* $\mathcal{M}'$ *with respect to* $\neg\phi$ *such that* $(\mathcal{M}'', \emptyset, \prec_{AD}) \models CHO_A(\phi)$.

Proof: Left-to-right follows directly from Theorem 13.1 and Theorem 13.3. To prove right-to-left, assume $(\mathcal{M}, \emptyset, \prec_{AD}) \models CHO_A(\phi) \wedge \neg INT_A(\phi)$. It follows that there exists $M \in Min(\mathcal{M}, \prec_{AD}) \cap Mod(\phi)$, and let $\mathcal{M}' = (\mathcal{M} \setminus Mod(\phi)) \cup \{M\}$. Then it is easy to see that for every partial restriction $\mathcal{M}''$ of $\mathcal{M}'$, $(\mathcal{M}'', \emptyset, \prec_{AD}) \models \neg CHO_A(\phi)$. □

Theorem 14.1

$$
\begin{aligned}
(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models\ & BEL_A(\neg hungry_4 \supset eat_3) \wedge \\
& INT_A(\neg hungry_4 \wedge \neg atCrowded_2) \wedge \\
& INT_A(\neg hungry_4) \wedge \\
& INT_A(eat_3) \wedge \\
& INT_A(eatSushi_3) \wedge \\
& GINT_A(atSushiBar_2) \wedge \\
& GINT_A(goSushiBar_1) \wedge \\
& GINT_A(\neg atCrowded_2).
\end{aligned}
$$

Proof: For all $M \in Mod(\mathcal{K})$ such that $M \models \neg hungry_4 \wedge \neg eat_3$, we can take $M' \in Mod(\mathcal{K})$ that satisfies $M' \models hungry_4$ and $M' \models p$ iff $M \models p$ for all the other atoms p. From a frame axiom 13, we have $M \prec_{\mathcal{P}} M'$. It follows that $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models BEL_A(\neg hungry_4 \supset eat_3)$. In the similar way, from 11 and 12, we have $(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models BEL_A(atSushiBar_2 \supset goSushiBar_1) \wedge BEL_A(atNoodleShop_2 \supset goNoodleShop_1)$.
To show that $\neg hungry_4 \wedge \neg atCrowded_2$, $\neg hungry_4$, eat_3, and $eatSushi_3$ are intentions, it is sufficient to show that $M \models eatSushi_3 \wedge \neg atCrowded_2$ for all $M \in Max(Max(Mod(\mathcal{K}), \prec_{\mathcal{P}}), \prec_{\mathcal{D}})$, and $N \models hungry_4$ for all $N \in Min(Max(Mod(\mathcal{K}), \prec_{\mathcal{P}}), \prec_{\mathcal{D}})$. This can be established as follows: First, there exist $M_1, M_2 \in Max(Mod(\mathcal{K}), \prec_{\mathcal{P}})$ such that $M_1 \models eatSushi_3 \wedge \neg eatSoba_3 \wedge \neg atCrowded_2$ and $M_2 \models \neg eatSushi_3 \wedge eatSoba_3 \wedge \neg eatUdon_3 \wedge \neg atCrowded_2$. From 14 and 15, all models of $\neg eat_3 \vee atCrowded_2$ are defeated (that is, less preferable with respect to $\prec_{\mathcal{D}}$) by M_1. From 16, all models of $\neg eatSushi_3 \wedge eatSoba_3$ are defeated by M_1. From 17, all models of $\neg eatSushi_3 \wedge \neg eatSoba_3 \wedge eatUdon_3$ are defeated by M_2, which is defeated by M_1 from 16. $Max(Mod(\mathcal{K}), \prec_{\mathcal{P}})$ includes a model of $hungry_4$, which is defeated by all models of $\neg hungry_4$ (14).
For generalized intentions, we can easily show that $atSushiBar_2$, $goSushiBar_1$ and $\neg atCrowded_2$ are not believed, and they are subgoals of $eatSushi_3$, $eatSushi_3$ and $\neg hungry_4 \wedge \neg atCrowded_2$, respectively. □

Theorem 14.2

$$\begin{aligned}(Mod(\mathcal{K}'), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models\ & INT_A(eatSoba_3) \wedge \\ & GINT_A(atNoodleShop_2) \wedge \\ & GINT_A(goNoodleShop_1) \wedge \\ & GINT_A(\neg atSushiBar_2) \wedge \\ & GINT_A(\neg goSushiBar_1).\end{aligned}$$

Proof: To show $(Mod(\mathcal{K}'), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models INT_A(eatSoba_3)$, it is sufficient to show that all models of $\neg eatSoba_3 \wedge eat_3$ are defeated by models of $eatSoba_3$. $Max(Mod(\mathcal{K}'), \prec_{\mathcal{P}})$ includes a model of $\neg eatSushi_3 \wedge eatSoba_3 \wedge \neg eatUdon_3 \wedge \neg atCrowded_2$, which defeats all models of $eatSushi_3 \wedge \neg eatSoba_3$ (15) and all models of $\neg eatSushi_3 \wedge \neg eatSoba_3 \wedge eatUdon_3$ (17).
For generalized intentions. $atNoodleShop_2$ and $goNoodleShop_1$ are subgoals of $eatSoba_3$, and $\neg atSushiBar_2$ and $\neg goSushiBar_1$ are subgoals of $\neg hungry_4 \wedge \neg atCrowded_2$, which is also an intention. They are obviously not believed. □

Theorem 14.3

$$(Mod(\mathcal{K}''), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models INT_A(eatUdon_3).$$

Proof: We only need the following observations:
$M \models \neg hungry_4 \wedge \neg atCrowded_2 \wedge \neg eatSushi_3 \wedge \neg eatSoba_3 \wedge eatUdon_3$ for all $M \in Max(Max(Mod(\mathcal{K}''), \prec_{\mathcal{P}}), \prec_{\mathcal{D}})$, and $N \models hungry_4$ for all $N \in Min(Max(Mod(\mathcal{K}''), \prec_{\mathcal{P}}), \prec_{\mathcal{D}})$. □

Theorem 14.4

$$\begin{aligned}(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models\ & BEL_A((C1 \supset INT_B(\neg hungryB_4)) \wedge \\ & (C1 \supset INT_B(eatB_3)) \wedge \\ & (C1 \wedge C2 \supset INT_B(eatSushiB_3)) \wedge \\ & (C1 \wedge C2 \wedge C3 \supset GINT_B(atSushiBarB_2)) \wedge \\ & (C1 \wedge C2 \wedge C3 \supset GINT_B(goSushiBarB_1)) \wedge \\ & (C1 \wedge C4 \supset INT_B(eatSobaB_3)) \wedge \\ & (C1 \wedge C4 \wedge C5 \supset GINT_B(atNoodleShopB_2)) \wedge \\ & (C1 \wedge C4 \wedge C5 \supset GINT_B(goNoodleShopB_1)))\end{aligned}$$

where

$$\begin{aligned}C1 &= \neg BEL_B(hungryB_4) \wedge \neg BEL_B(\neg hungryB_4), \\ C2 &= \neg BEL_B(eatSushiB_3 \wedge \neg eatSobaB_3 \supset atCrowdedB_2) \wedge \\ &\quad \neg BEL_B(\neg eatSushiB_3 \wedge eatSobaB_3 \wedge \neg eatUdonB_3 \supset atCrowdedB_2), \\ C3 &= \neg BEL_B(atSushiBarB_2), \\ C4 &= BEL_B(eatSushiB_3 \wedge \neg eatSobaB_3 \supset atCrowdedB_2) \wedge \\ &\quad \neg BEL_B(\neg eatSushiB_3 \wedge eatSobaB_3 \wedge \neg eatUdonB_3 \supset atCrowdedB_2), \\ C5 &= \neg BEL_B(atNoodleShopB_2).\end{aligned}$$

Proof: Let $M = (W, \prec_{BP}, \prec_{BD}, w_0)$ be an element of $Max(Mod(\mathcal{K}), \prec_{\mathcal{P}})$ that satisfies Condition $C1$. $Max(W, \prec_{BP})$ includes both a model of $hungryB_4$ and a model of $\neg hungryB_4$, and from Sentence 14, all models of $hungryB_4$ are defeated by models of $\neg hungryB_4$. It follows that $M \models INT_B(\neg hungryB_4)$, and since $\neg hungryB_4$ and $eatB_3$ are equivalent in $Max(Mod(\mathcal{K}), \prec_{\mathcal{P}})$, $M \models INT_B(eatB_3)$. Suppose M satisfies $C2$. To show $M \models INT_B(eatSushiB_3)$, it is sufficient to show that all models of $\neg eatSushiB_3 \wedge eatB_3$ are defeated by models of $eatSushiB_3$. From 15 and 17, all models of $\neg eatSushiB_3 \wedge eatSobaB_3$ are defeated by models of $eatSushiB_3 \wedge \neg eatSobaB_3 \wedge \neg atCrowdedB_2$. From 15 and 22, all models of $\neg eatSushiB_3 \wedge \neg eatSobaB_3 \wedge eatUdonB_3$ are defeated by models of $\neg eatSushiB_3 \wedge eatSobaB_3 \wedge \neg eatUdonB_3 \wedge \neg atCrowdedB_2$.
For generalized intentions. Since $atSushiBarB_2$ and $goSushiBarB_1$ are subgoals of $eatSushiB_3$, if they are not believed, they are generalized intentions.
Next, suppose M satisfies $C4$. To show $M \models INT_B(eatSobaB_3)$, it is sufficient to show that all models of $\neg eatSobaB_3 \wedge eatB_3$ are defeated by models of $eatSobaB_3$. From $C4$, $Max(W, \prec_{BP})$ includes a model of $\neg eatSushiB_3 \wedge eatSobaB_3 \wedge \neg eatUdonB_3 \wedge \neg atCrowdedB_2$, which defeats all models of $eatSushiB_3 \wedge \neg eatSobaB_3$ (15) and all models of $\neg eatSushiB_3 \wedge \neg eatSobaB_3 \wedge eatUdonB_3$ (15 and 22).
For generalized intentions. Since $atNoodleShopB_2$ and $goNoodleShopB_1$ are subgoals of $eatSobaB_3$, if they are not believed, they are generalized intentions. □

Theorem 14.5

$$\begin{aligned}(Mod(\mathcal{K}), \prec_{\mathcal{P}'}, \prec_{\mathcal{D}}) \models\ & BEL_A(INT_B(\neg hungryB_4)) \wedge \\ & INT_B(eatB_3) \wedge \\ & INT_B(eatSushiB_3) \wedge \\ & GINT_B(atSushiBarB_2) \wedge \\ & GINT_B(goSushiBarB_1)).\end{aligned}$$

Proof: In view of the last theorem, we only need to show that $C1$, $C2$ and $C3$ are satisfied in all $M \in Max(Mod(\mathcal{K}), \prec_{\mathcal{P}'})$. This is established by the following easy observation: for all $M \in Max(Mod(\mathcal{K}), \prec_{\mathcal{P}'})$, $M \models BEL_B(\alpha)$ if and only if α is a logical consequence of B's beliefs specified by Sentences 1–13. □

Theorem 14.6

$$\begin{aligned}(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models\ & BEL_A(knowGateB_2 \supset informGateA_1) \wedge \\ & BEL_A(atMuseumB_4 \supset gotoMuseumB_3) \wedge \\ & INT_A(watchParadeB_6) \wedge \\ & INT_A(enterMuseumB_5) \wedge \\ & GINT_A(atMuseumB_4) \wedge \\ & GINT_A(gotoMuseumB_3) \wedge\end{aligned}$$

$$GINT_A(knowGateB_2) \wedge$$
$$GINT_A(informGateA_1) \wedge$$
$$GINT_A(openMuseum_4).$$

Proof: This theorem is easily proved in the similar way to Theorem 14.1. □

Theorem 14.7

$$(Mod(\mathcal{K}'), \prec_P, \prec_D) \models \neg INT_A(enterMuseumB_5).$$
$$(Mod(\mathcal{K}''), \prec_P, \prec_D) \models \neg INT_A(enterMuseumB_5).$$

Proof: For all $M \in Max(Max(Mod(\mathcal{K}'), \prec_P), \prec_D)$, $M \models \neg openMuseum_4$, and thus $M \models \neg enterMuseumB_5$. For all $M' \in Max(Max(Mod(\mathcal{K}''), \prec_P), \prec_D)$, $M' \models watchParadeB_6$ (9), and thus $M' \models \neg enterMuseumB_5$. □

Theorem 14.8

$$(Mod(\mathcal{K}), \prec_P, \prec_D) \models$$
$$BEL_A(\quad BEL_B(atMuseumB_3 \supset gobyBusB_2 \vee goOnFootB_2) \wedge$$
$$(\neg BEL_B(gobyBusB_2 \supset \neg goOnFootB_2) \supset INT_B(atMuseumB_3)) \wedge$$
$$D\text{-}PREF_B^{eq}(gobyBusB_2, goOnFootB_2)).$$

Proof: Let $M = (W, \prec_{BP}, \prec_{BD}, w_0)$ be an element of $Max(Mod(\mathcal{K}), \prec_P)$. First, we want to show that $Max(W, \prec_{BP})$ does not include a model of $atMuseumB_3 \wedge \neg gobyBusB_2 \wedge \neg goOnFootB_2$. If it includes such a model, that model defeats all models of $gobyBusB_2$ (2, 4), but this conflicts with 5.
Next, suppose $M \models \neg BEL_B(gobyBusB_2 \supset \neg goOnFootB_2)$. This means that there is $w \in Max(W, \prec_{BP})$ such that $w \models gobyBusB_2 \wedge goOnFootB_2$. Since w is a model of an intention $gobyBusB_2$ and defeated by all the other models of $atMuseumB_3$, it must defeat the only model of $\neg atMuseumB_3$. It follows that $M \models INT_B(atMuseumB_3)$. Finally, we have $M \models D\text{-}PREF_B^{eq}(gobyBusB_2, goOnFootB_2)$, because if there is $w' \in Max(W, \prec_{BP})$ such that $w' \models \neg gobyBusB_2 \wedge goOnFootB_2$, from 2 and 5, w' is defeated by a model of $gobyBusB_2 \wedge \neg goOnFootB_2$. □

Theorem 14.9

$$(Mod(\mathcal{K}), \prec_P, \prec_D) \models BEL_A(INT_B(gobyBusB_2) \wedge \neg INT_B(knowGateB_1)).$$

Proof: $knowGateB_1$ is a subgoal of $knowGateB_1$, $gobyBusB_2$, or $knowGateB_1 \wedge \neg gobyBusB_2$. Therefore, for all $M \in Max(Mod(\mathcal{K}), \prec_P)$, we have $M \models \neg INT_B(knowGateB_1)$ from 3, and then we have $M \models INT_B(gobyBusB_2)$ from 4. □

Theorem 14.10

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}'}, \prec_{\mathcal{D}}) \models BEL_A(INT_B(gobyBusB_2)).$$

Proof: This theorem follows obviously from 2 and $\mathcal{P}'$. □

Theorem 14.11

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models BEL_A(INT_B(atMuseumB_3 \wedge spendMoneyB_3) \wedge INT_B(atMuseumB_3)).$$

Proof: Let $M = (W, \prec_{BP}, \prec_{BD}, w_0)$ be an element of $Max(Mod(\mathcal{K}), \prec_{\mathcal{P}})$. To prove the theorem, it is sufficient to show that $atMuseumB_3 \wedge spendMoneyB_3$, $atMuseumB_3$ and $gobyBusB_2$ are equivalent in $Max(W, \prec_{BP})$. To show this, we want to show that $Max(W, \prec_{BP})$ includes neither a model of $\neg gobyBusB_2 \wedge atMuseumB_3 \wedge spendMoneyB_3$ nor a model of $\neg gobyBusB_2 \wedge atMuseumB_3 \wedge \neg spendMoneyB_3$. If it includes either of these models, the only model of $gobyBusB_2$ is defeated by that model (2, 3), but this conflicts with 4. □

Theorem 14.12

$$(Mod(\mathcal{K}'), \prec_{\mathcal{P}'}, \prec_{\mathcal{D}'}) \models BEL_A(INT_B(atMuseumB_3 \wedge spendMoneyB_3) \wedge INT_B(atMuseumB_3)).$$

Proof: The first conclusion is obtained in the same way as the last theorem, and then the second conclusion follows obviously from 5 and 6. □

Theorem 14.13

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}}, \prec_{\mathcal{D}}) \models BEL_A((INT_B(huntB_3) \wedge INT_B(cashCheckB_3)) \vee INT_B(robBankB_3)).$$

Proof: This theorem follows directly from Sentences 1,2,3 and 4. □

Theorem 14.14

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}'}, \prec_{\mathcal{D}}) \models BEL_A(INT_B(robBankB_3)).$$

Proof: This theorem follows obviously from the last theorem and $\mathcal{P}'$. □

Theorem 14.15

$$(Mod(\mathcal{K}), \prec_{\mathcal{P}''}, \prec_{\mathcal{D}}) \models BEL_A(INT_B(huntB_3) \wedge (INT_B(cashCheckB_3) \vee INT_B(robBankB_3)).$$

Proof: Since $\neg INT_B(huntB_3)$ implies $INT_B(robBankB_3)$ in $Max(Mod(\mathcal{K}), \prec_{\mathcal{P}''})$, all models of $\neg INT_B(huntB_3)$ are defeated by models of $INT_B(huntB_3) \wedge \neg INT_B(robBankB_3)$. In view of Theorem 14.13, this proves the theorem. □

Bibliography

[1] J. F. Allen. Recognizing intentions from natural language utterances. In Michael Brady and Robert C. Berwick, editors, *Computational Models of Discourse*, pages 107–166. MIT Press, 1983.

[2] J. F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154, 1984.

[3] J. F. Allen and J. A. Kooman. Planning using a temporal world model. In *Proceedings of IJCAI–83*, pages 741–747, 1983.

[4] D. E. Appelt. *Planning English Sentences.* Cambridge University Press, 1985.

[5] Michael E. Bratman. *Intention, Plans and Practical Reason.* Harvard University Press, 1987.

[6] Sandra Carberry. *Plan Recognition in Natural Language Dialogue.* MIT Press, 1990.

[7] D. Chapman. Planning for conjunctive goals. *Artificial Intelligence*, 32:333–377, 1987.

[8] E. Charniak. A neat theory of marker passing. In *Proceedings of AAAI–86*, pages 584–588, 1986.

[9] Brian F. Chellas. *Modal Logic: An Introduction.* Cambridge University Press, 1980.

[10] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[11] Philip R. Cohen and Hector J. Levesque. Rational interaction as the basis for communication. In Philip R. Cohen, Jerry L. Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 221–255. MIT Press, 1990.

[12] K. Devlin. *Logic and Information.* Cambridge University Press, 1991.

[13] R. E. Fikes and N. J. Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2:189–208, 1971.

[14] Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* MIT Press, 1988.

[15] Matthew L. Ginsberg. Possible worlds planning. In Michael P. Georgeff and Amy L. Lansky, editors, *Reasoning about Actions and Plans, Proceedings of the 1986 Workshop*, pages 213–243. Morgan Kauffman, 1987.

[16] J. Y. Halpern and Y. Moses. A guide to the modal logics of knowledge and belief: Preliminary draft. In *Proceedings of IJCAI–85*, pages 480–490, 1985.

[17] D. Harel. *First–Order Dynamic Logic, Lecture Notes in Computer Science 68.* Springer-Verlag, 1979.

[18] K. Hasida and S. Isizaki. Dependency propagation: A unified theory of sentence comprehension and generation. In *Proceedings of IJCAI–87*, pages 664–670, 1987.

[19] G. Hirst. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press, 1987.

[20] Jerry R. Hobbs, Mark E. Stickel, Douglas E. Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63:69–142, 1993.

[21] Hitoshi Iida and Hidekazu Arita. Natural language dialogue understanding on a four-layer plan recognition model. *Transactions of IPSJ*, 31(6):810–821, 1990.

[22] J. K. Kaplan. Cooperative responses from a portable natural language database query system. In Michael Brady and Robert C. Berwick, editors, *Computational Models of Discourse.* MIT Press, 1983.

[23] Henry A. Kautz. A circumscriptive theory of plan recognition. In Philip R. Cohen, Jerry L. Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 105–133. MIT Press, 1990.

[24] Kurt Konolige and Martha E. Pollack. Ascribing plans to agents: Preliminary report. In *Proceedings of IJCAI–89*, pages 924–930, 1989.

[25] Kurt Konolige and Martha E. Pollack. A representationalist theory of intention. In *Proceedings of IJCAI–93*, pages 390–395, 1993.

[26] J. Savage Leonard. *The Foundations of Statistics.* Wiley, 1954.

[27] Hector J. Levesque, Philip R. Cohen, and José H. T. Nunes. On acting together. In *Proceedings of AAAI–90*, pages 94–99, 1990.

[28] J. McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28:89–116, 1986.

[29] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502, 1969.

[30] C. Mellish. *Computer Interpretation of Natural Language Descriptions.* Ellis Horwood, 1985.

[31] H. Nakashima, S. Peters, and H. Schütze. Communication and inference through situations. In *Proceedings of IJCAI–91*, pages 76–81, 1991.

[32] Martha E. Pollack. Plans as complex mental attitudes. In Philip R. Cohen, Jerry L. Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 77–103. MIT Press, 1990.

[33] Anand S. Rao and Michael P. Georgeff. Asymmetry thesis and side-effect problems in linear-time and branching-time intention logics. In *Proceedings of IJCAI–91*, pages 498–504, 1991.

[34] P. Rosenbloom, J. Laird, and A. Newell. Meta–levels in soar. In P. Maes and D. Nardi, editors, *Meta–Level Architecture and Reflection*, pages 227–239. Elsevier Science Publishers, 1988.

[35] E. Sacerdoti. *A Structure for Plans and Behavior.* North Holland, 1977.

[36] M. D. Sadek. A study in the logic of intention. In *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning*, pages 462–473, 1992.

[37] Ken Satoh. A logical formalization of preference-based reasoning by interpretation ordering. Doctoral thesis, Department of Information Science, Faculty of Science, University of Tokyo, 1992.

[38] R. C. Schank and R. P. Abelson. *Scripts, Plans, Goals and Understanding.* John Wiley and Sons, 1977.

[39] H. Schütze. *The PROSIT Language v0.4*, 1991.

[40] John R. Searle. *Speech Acts.* Cambridge University Press, 1969.

[41] John R. Searle and Daniel Vanderveken. *Foundations of Illocutionary Logic.* Cambridge University Press, 1985.

[42] Glenn Shafer. Savage revisited. *Statistical Science*, 1(4):463–485, 1986.

[43] Yoav Shoham. *Reasoning about Change.* MIT Press, 1988.

[44] Yoav Shoham and Steve B. Cousins. Logics of mental attitudes in AI: a very preliminary survey. In Gerhard Lakemeyer and Bernhard Nebel, editors, *Foundations of Knowledge Representation and Reasoning, Lecture Notes in Artificial Intelligence 810*, pages 296–309. Springer-Verlag, 1994.

[45] E. H. Shortliffe. *Computer-Based Medical Consultations: MYCIN.* Elsevier, 1976.

[46] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, volume 2, pages 105–134. Kluwer Academic Publishers, Dordrecht, 1987.

[47] Toru Sugimoto and Akinori Yonezawa. Multiple world representation of mental states for dialogue understanding (in japanese). In *91-NL-83-5*, pages 31–38. Information Processing Society of Japan, May 1991.

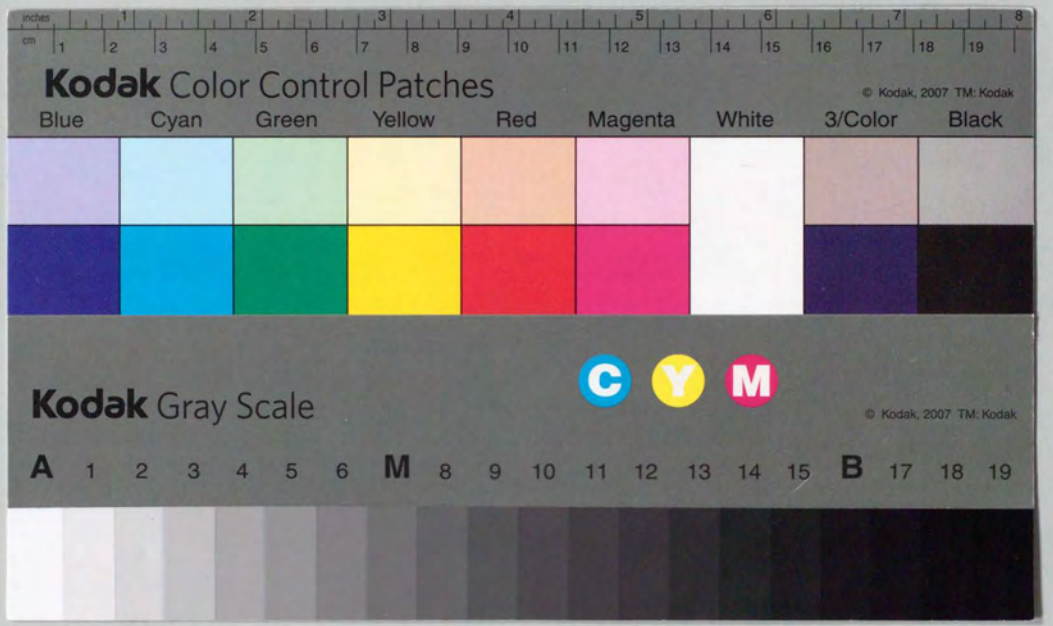
[48] Toru Sugimoto and Akinori Yonezawa. Multiple world representation of mental states for dialogue processing. *IEICE Transactions on Information and Systems*, E77-D(2):192–208, February 1994.

[49] Toru Sugimoto and Akinori Yonezawa. A preference-based theory of intention. Technical Report 94–4, Department of Information Science, Faculty of Science, University of Tokyo, February 1994.

[50] Michael P. Wellman. Qualitative probabilistic networks for planning under uncertainty. In Glenn Shafer and Judea Pearl, editors, *Readings in Uncertain Reasoning*, pages 711–722. Morgan Kaufmann Publishers, 1990.

[51] Emil Weydert. Hyperrational conditionals: Monotonic reasoning about nested default conditionals. In Gerhard Lakemeyer and Bernhard Nebel, editors, *Foundations of Knowledge Representation and Reasoning, Lecture Notes in Artificial Intelligence 810*, pages 310–332. Springer-Verlag, 1994.

[52] Y. Wilks. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6:53–74, 1975.


inches
cm
Kodak Color Control Patches
© Kodak, 2007 TM: Kodak
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
Kodak Gray Scale
C
Y
M
© Kodak, 2007 TM: Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19