On Extending BDI Logics

Vineet Chand Padmanabhan Nair

Knowledge Representation & Reasoning Unit (KRRU)
Griffith University, Gold Coast Campus
Queensland, Australia

On Extending BDI Logics

A thesis presented

by

Vineet Chand Padmanabhan Nair

to

The School of Information Technology
in fulfillment of the requirements
for the degree of
Doctor of Philosophy

Faculty of Engineering and Information Technology
Griffith University, Queensland
Australia

4^{th} March 2003

For Amma and Acchan

Abstract

In this thesis we extend BDI logics, which are normal multimodal logics with an arbitrary set of normal modal operators, from three different perspectives. Firstly, based on some recent developments in modal logic, we examine BDI logics from a *combining* logic perspective and apply combination techniques like *fibring/dovetailing* for explaining them. The second perspective is to extend the underlying logics so as to include action constructs in an explicit way based on some recent action-related theories. The third perspective is to adopt a non-monotonic logic like *defeasible logic* to reason about intentions in BDI. As such, the research captured in this thesis is theoretical in nature and situated at the crossroads of various disciplines relevant to Artificial Intelligence (AI). More specifically this thesis makes the following contributions:

- **Combining BDI Logics through *fibring/dovetailing* [61]**: BDI systems modeling rational agents have a combined system of logics of belief, time and intention which in turn are basically combinations of well understood modal logics. The idea behind combining logics is to develop *general techniques* that allow to produce combinations of *existing* and *well understood* logics. To this end we adopt Gabbay's [42] fibring/dovetailing technique to provide a general framework for the combinations of BDI logics. We show that the existing BDI framework is a dovetailed system. Further we give conditions on the fibring function to accommodate interaction axioms of the type $G^{k,l,m,n}$ $(\Diamond^k\Box^l\varphi \Rightarrow \Box^m\Diamond^n\varphi)$ based on Catach's multimodal semantics. This is a major result when compared with other combining techniques like *fusion* which fails to accommodate axioms of the above type.

- **Extending the BDI framework to accommodate *Composite Actions* [98]**: Taking motivation from a recent work on BDI theory [97], we incorporate the notion of *composite actions*, $\pi_1;\pi_2$ (interpreted as π_1 *followed by* π_2), to the existing BDI framework. To this

end we introduce two new constructs *Result* and *Opportunity* which helps in reasoning about the actual execution of such actions. We give a set of axioms that can accommodate the new constructs and analyse the set of *commitment* axioms as given in the original work in the background of the new framework.

- **Intention reasoning as Defeasible reasoning [59, 60]**: We argue for a non-monotonic logic of intention in BDI as opposed to the usual normal modal logic one. Our argument is based on Bratman's *policy-based intention* [12]. We show that policy-based intention has a *defeasible/non-monotonic* nature and hence the traditional *normal modal logic* approach to reason about such intentions fails. We give a formalisation of policy-based intention in the background of *defeasible logic*. The problem of logical omniscience which usually accompanies normal modal logics is avoided to a great extend through such an approach.

Contents

List of Figures **v**

List of Tables **vi**

1 Introduction **1**
1.1 Representation & Reasoning with BDI 1
1.2 What This Thesis is About 5
1.3 How We Go About Solving It (Tools Used) 6
1.4 Plan Of This Thesis 7
1.5 Bibliographic Note 8

2 Intensional Logics and BDI **10**
2.1 Modal Logic (The Basic Case) 10
2.2 Multi-Modal Logics 25
2.3 Combining Logics 38
2.4 Summary and Discussion 50

3 Fibring Semantics for BDI Logics **52**
3.1 Background and motivations 52
3.2 The Problem 53
3.3 Fibring of BDI Logics 55
3.4 Semantics for Mental States 58
3.5 Conditions on the Fibring Function 62
3.6 Summary and Discussion 71

4 Extending BDI with Composite Actions **73**
4.1 Background and Motivations 73
4.2 Modal Logics of Action and Ability 74
4.3 Integrating Results 83
4.4 Integrating Opportunity 86
4.5 Opportunity + Results 88

4.6 Commitment Axioms Revisited 89
4.7 Summary and Discussion 90

5 Intention Reasoning as Defeasible Reasoning 92
5.1 Background and Motivations 92
5.2 Logics of Intention: An overview 94
5.3 The Case for Non-Monotonic Reasoning 101
5.4 Logical Omniscience and Non-Monotonicity 103
5.5 Overview of Defeasible Logic 104
5.6 Defeasible Logic for Intentions 108
5.7 Revisiting Intention Reconsideration and Commitment Strategies 115
5.8 Summary and Discussion 116

6 Conclusions and Future Work 118
6.1 Summary 118
6.2 Future Work 120

Bibliography 123

List of Figures

2.1 A Kripke model (A) and Kripke frame (B) 13
2.2 A Kripke frame satisfying the formula $\Box \mathfrak{p}_0 \Rightarrow \mathfrak{p}_0$ (1A) and the associated Kripke Model (1B) 17
2.3 Realism possible worlds structure ($\mathsf{I} \subseteq \mathsf{G} \subseteq \mathsf{B}$) 34
2.4 Strong realism possible worlds structure ($\mathsf{B}^{\omega} \subseteq_{sup} \mathsf{G}^{\omega} \subseteq_{sup} \mathsf{I}^{\omega}$) . . 35
2.5 Weak realism possible worlds structure 38
2.6 An example of Fibring 45

3.1 a, b, c, d incestuality property 61
3.2 Incestuality with respect to cross-modality relations 68
3.3 Inclusion morphism $\oplus$ fibring $\oplus$ incestuality 69

5.1 An example of belief, desire and intention revision 98

List of Tables

2.1 Some important logics 16
2.2 The property of R corresponding to some formulas 18
2.3 Completeness table for some of the basic Logics 23
2.4 Examples of Multi-Modal Logics 26

3.1 Some well known axiom schemas included by the incestual axioms 60

4.1 Axioms related to the I-system of Rao and IC-system of Padgham 83

5.1 Structural relationship among modalities in G&R formalism . . 99

Acknowledgments

With a sense of privilege I thank the co-architects of this dissertation:-

Prof. Abdul Sattar (my supervisor) for locating me in the outskirts of *The land of Kohinoor* (the city of Hyderabad) and bringing me to the *Sunshine state* (the city of Brisbane) for doing research in Artificial Intelligence. I am indebted to him for putting his trust on someone he barely knew. Thanks a lot Abdul.

Dr. Guido Governatori (my co-supervisor) who taught me how to do logic. It is quite difficult to recollect the exact day I started liking logic but it was definitely after my acquaintance with Guido. The seeds must have been paved much earlier but it was Guido who showed me the way. I am really proud of getting him as a supervisor.

Prof. Arun Kumar Pujari (my external supervisor) for initiating the idea of inter-disciplinary research while I was a student at Hyderabad Central University. I should also thank Prof. Chinmoy Goswamy for the same.

Prof. Dov Gabbay (King's College, London) and Prof. Norman Foo (UNSW, Sydney) for their valuable comments as external examiners. Norman's precision in pointing out the latex errors improved the final version to a great extend. Prof. Rodney Topor for his role as chief of examiners.

School of information technology (INT-Gold Coast campus) and school of communication and information technology (CIT-Nathan campus) for providing the financial support during my stint as a Ph.D student. Special thanks to school of INT for providing a stimulating and pleasant working environment. Visiting conferences can be expensive. I am therefore very grateful to school of INT for providing support of the expenses associated with attending conferences/workshop's. Thanks are also due to JELIA-02 organising committee for supporting me with a student scholarship and the

financial support from ARC large grant on *Dynamics of mental states of rational agents*.

Members of the K.R.R.U lab for their support and enthusiasm. I cannot forget the help I got from Richard, Steve and Shiv during my initial days as a doctoral student. Richard for helping me out with latex and Steve/Shiv for introducing me to the "Linux" world. Lin Zhou was always there as the next door neighbour and Valnir for being my soccer "Guru". Thanks are also due to the new comer's. Stuart for his organisational capacity and Owen/Tim for their placidity. Justin was always there as a friend. Anbu's occassional visits to have coffee gave time to relax. Last but not the least John Thornton who was always interested in the *big questions*.

On a personal level I would like to thank Jeff (my flat mate) for putting up with me during my hard times. *Friday night at Jeff's* is something worth for a life time. Thanks in ton mate. Sathya and Kavita for being part of the colourful crowd at 20 Meldrum during my initial days. Barbara and Dorothy for making me feel at home. Rino and Tracy for the neighborhood entertainment. Jessica for her help in proof reading and Matt for *matteism*. Suli and Jessy for giving me shelter during the final days and Sasi for being part of the 7-eleven debacle.

Members, (old as well as new), of the *mallu-mafia* who always stood by me during my days in Hyderabad. Special thanks to the *core* comprising of Finney, Binu, Sreeku, Pengal, Saji and Bobby. Ettan and Shal, Deepthi and Prasanth and of course Binu Zach needs special mention.

Last but not the least my family to whom I owe this work. For my father and mother who always supported my studies. For my brothers who were always willing to listen and do the needful. I wouldn't have reached this stage without their support.

Chapter 1

Introduction

If you understand, things are just as they are, if you do not understand, things are just as they are.

Zen proverb

In this chapter we briefly explain the *belief, desire, intention* (BDI) strategy and its underlying logical framework. We then examine how and why the existing logical framework could be extended. It is argued that modal logic provides a good formal tool to model BDI-like systems. Finally, we conclude the chapter with an overview of the structure of the work presented in this thesis.

1.1 Representation & Reasoning with BDI

1.1.1 The BDI Strategy

There are many strategies for representing and reasoning about the problem-solving mechanisms involved in artificial entities like a robot/computer system. For example, consider the physical strategy (or *physical stance*) through which, one predicts the behavior of a robot by determining its physical constitution (all the way down to the micro-physical level) using the knowledge of the laws of physics and thereby predicts the outcome for any input. Such a strategy was used by Laplace[1] for predicting the entire

[1]French mathematician (1749-1827) known for Laplace's equation, which is one of the important partial differential equations of physics.

future of everything in the universe, but the difficulty is that the strategy is not always practically viable. Then there is the *design stance* which helps in predicting that a particular entity will behave *as it is designed to behave* under various circumstances. For instance, most computer users have the faintest idea about the hardware components in their system but if they have a good idea of the operating system (the way the computer is designed to behave) they can predict its behavior with great accuracy and reliability. But as the application domains become increasingly large and complex, as in the design of a self contained problem solving system capable of *autonomous, reactive, pro-active, social behavior* (often termed as *agents*), at times one is urged to adopt a stance through which the problem and its solution could be expressed more naturally. In this thesis we deal with such a strategy called the *intentional stance* to explain the rational behavior involved in BDI-like agents. The strategy is stated as follows:

> *The intentional strategy consists of treating the object whose behavior you want to predict as a rational agent with beliefs and desires and other mental states exhibiting intentionality.*
> Daniel C. Dennet

Based on the above paradigm and following the philosophical work of Bratman [12], Dennet [29] and the more formalised version of Cohen and Levesque [25], Rao and Georgeff developed their own formalism for modeling rational agents [100, 102, 103, 107, 106], which they called BDI-*agent framework*. BDI agents are systems[2] that are situated in a changing environment, receive continuous perceptual input, and take actions to affect their environment, all based on their internal mental states. The BDI model focuses on three components of an agent: its beliefs, its desires (goals)[3] and its intentions, representing respectively, the information, the motivational and the decision states of the agent. The notion of intention has equal status with the notions of belief and desire, and cannot be reduced to these concepts. This allows different types of rational agents to be modeled by imposing certain conditions on the persistence of an agent's beliefs, goals and intentions.

Since the seminal work of Rao and Georgeff [107], the last decade has witnessed a flowering interest in BDI-based agent research both on a theoretical and practical level. The theoretical level, often referred to as the *logical level*, is concerned with specifying the properties of a BDI agent

[2]Systems could be either defined as a complex of language, models and interpretations from a logical point of view or a computer system capable of rational behavior.

[3]In this thesis we do not differentiate between desires and goals.

and their formal representation in some logical language. From a system-building perspective this level is called the *specification* phase. The logical level tries to answer questions related to the various components of a BDI agent's cognitive makeup and how to derive rational behavior from such a setup. It is also concerned with providing explanations as to the relationships between the different cognitive components, the need for logic based formalisms etc.

On the other hand, the practical side is divided into two levels called the *architectural* level and the *language* level. At the architecture level, issues related to the construction of computer systems that satisfy the properties specified at the logical level is of prime concern, and is usually called the *design phase*. The BDI architecture is often *deliberative* in nature in the sense that it contains an explicitly represented, symbolic model of the world and the decisions are made via logical reasoning. This is in contrast to the *reactive* architectures [13], which do not include any kind of central symbolic model and does not use complex symbolic reasoning. In the later works an *interactive* architecture for BDI is also considered [37, 63]. Finally, the language level (*implementation phase*), is concerned with developing an interpreted programming language which includes at least some structure corresponding to a BDI agent as is done in [126, 102, 15, 11]. There are also a number of BDI implementations which are used successfully in critical application domains as well as for implementing business, operational and advanced reasoning procedures (such as PRS [34], DMARS, JAM [68] and JACK [16]). Further details are listed at (http://www.agent-software.com/) and for a more comprehensive account on agents see (http://www.agentbuilder.com).

Recent publications [94, 11, 97] show that advancement in BDI research is happening at all three levels rapidly. This thesis offers no contribution to the practical side but tries to enhance our understanding of the logical level.

1.1.2 BDI Logics

BDI logics are multi-modal logics with three families of modal operators BEL_i, DES_i and INT_i, where $i \in A$ (agents). Multi-modal logics generalise modal logics allowing more than one modal operator to appear in formulae. In particular, a modal operator is named by means of a label, for instance $\Box_i$ which identifies it. Hence a formula like $\Box_i\varphi$ could be interpreted as φ *is believed by the agent* i or φ *is a goal for agent* i or φ *is true after executing the action* i, representing respectively the belief, goal and action constructs of an agent. In addition, the BDI formalism extends a branching

time logic called CTL^* (Computation Tree Logic) [33] to first-order logic along with the modalities for belief, goal and intention[4]. Beliefs, desires and intentions are modeled as sets of belief-, goal-, and intention-accessible worlds[5] associated to an agent in each *situation* (a particular time point in a particular world is called a *situation*). The semantics is given in terms of standard Kripke frames. Three families of accessibility relations are defined, one each for belief, $\{\text{BEL}_i\}$, intention, $\{\text{INT}_i\}$ and desire, $\{\text{DES}_i\}$. The belief accessibility relation is *serial, Euclidean* and *transitive* whereas for intention and goal the accessibility relation is *serial*. The properties which characterise the modal operators BEL, GOAL and INT are expressed in the form of axioms. One such axiom that is common to all the operators is called axiom **K** and is given as follows

$$\Box_i(\varphi \Rightarrow \psi) \Rightarrow (\Box_i\varphi \Rightarrow \Box_i\psi) \tag{1.1}$$

The modal systems that are characterised by the above axiom are called *normal modal logics* and the modal operators involved *normal modal operators*. Hence BEL, GOAL and INT are known as normal modal operators. If we interpret $\Box_i$ as BEL, then axiom (1.1) conveys the meaning that an agent *believes all the consequences of its beliefs*. In a similar manner other axioms satisfying specific properties could be given for each modal operator. One key assumption made in the BDI framework is that of *intentions being stronger than desires and desires being stronger than beliefs*. These constraints too are given in the form of axioms as shown below

$$\text{INT}(\varphi) \Rightarrow \text{GOAL}(\varphi) \tag{1.2}$$

$$\text{GOAL}(\varphi) \Rightarrow \text{BEL}(\varphi) \tag{1.3}$$

These axioms capture the rich inter-relationships between beliefs, desires and intentions and are often termed *interaction axioms*. The **K**-axiom (1.1) is also of this type but the difference is that axioms (1.2) and (1.3) are *non-homogeneous* (i.e., every modal operator is not restricted to the same system. For instance the underlying axiom systems for GOAL is **K** and **D** of modal logic whereas that of BEL is **KD45**. In this thesis we are interested in interaction axioms of the type (1.2) and (1.3) but having a general form like

$$\Box_{i_1}\Box_{i_2}\ldots\Box_{i_n}\varphi \Rightarrow \Box'_{i_1}\Box'_{i_2}\ldots\Box'_{i_m}\varphi \ (n > 0, m \geq 0)$$

and the modal systems charactrised by such general schemas.

[4]We do not make any reference to temporal constructs in this thesis.

[5]The notion is related to *possible worlds* which we explain in Chapter 2

In addition to BEL, GOAL and INT, other notions such as commitments, capabilities, know-how, etc. have been investigated within the framework of BDI. Sophisticated multi-modal, temporal, action, and dynamic logics have been used to formalize these notions, see among others, [97, 123, 117, 132, 111, 38]. Hence, as stated in [19], the main feature of multi-modal systems is their ability to express *complex modalities*, obtained by composing modal operators of different types. Such systems allow one to design agent situations where the agents have different ways of reasoning and different ways of interacting between them thereby allowing simultaneous study of several modal aspects (e.g., belief and knowledge, belief and action etc.)

1.2 What This Thesis is About

In this thesis we extend the underlying logics in BDI-systems from three different perspectives. Firstly, based on some recent developments in modal logic, we examine BDI logics from a *combining* logic perspective and apply combination techniques like *fibring/dovetailing* for explaining them. The second perspective is to extend the underlying logics so as to include action constructs in an explicit way based on some recent action-related theories. The third perspective is to adopt a non-monotonic logic like *defeasible logic* to reason about intentions in BDI. As such, the research captured in this thesis is theoretical in nature and situated at the crossroads of various disciplines relevant to Artificial Intelligence (AI). More specifically this thesis makes the following contributions:

- **Combining BDI logics through *fibring/dovetailing* [61]**: BDI systems modeling rational agents have a combined system of logics of belief, time and intention which in turn are basically combinations of well understood modal logics. The idea behind combining logics is to develop *general techniques* that allow to produce combinations of *existing* and *well understood* logics. To this end we adopt Gabbay's [42] fibring/dovetailing technique to provide a general framework for combining BDI logics. We show that the existing BDI framework is a dovetailed system. Such general techniques help in formalising complex systems like BDI in a systematic way i.e., such a general methodology would permit a modular treatment of the modal components, whereby each component is analysed and developed on its own with the most appropriate methodology for it and reused in the combination. Further we give conditions on the fibring function to accommodate interaction axioms of the type $G^{k,l,m,n}$ $(\Diamond^k\Box^l(\varphi) \Rightarrow \Box^m\Diamond^n\varphi)$

based on Catach's multimodal semantics. This is a major result when compared with other combining techniques like *fusion* which fails to accommodate axioms of the above type.

- **Extending the BDI framework to accommodate *composite actions* [98]**: Taking motivation from a recent work on BDI theory [97], we incorporate the notion of *composite actions*, $\pi_1; \pi_2$ (interpreted as π_1 *followed by* π_2), to the existing BDI framework. To this end we introduce two new constructs *Result* and *Opportunity* which help in reasoning about the actual execution of such actions. We give a set of axioms that accommodates the new constructs.

- **Intention reasoning as defeasible reasoning [59, 60]**: We argue for a non-monotonic logic of intention in BDI as opposed to the usual normal modal logic one. Our argument is based on Bratman's *policy-based intention* [12]. We show that policy-based intention has a *defeasible/non-monotonic* nature and hence the traditional *normal modal logic* approach to reason about such intentions fail. We give a formalisation of policy-based intention in the background of *defeasible logic*. The problem of logical omniscience which usually accompanies normal modal logics is avoided to a great extent through such an approach.

1.3 How We Go About Solving It (Tools Used)

The main formal tool used is Modal Logic [69, 70, 22, 21, 76, 10], as it is widely known these days that they are suitable for structuring and representing cognitive constructs like knowledge and belief [73, 51, 64] as well as other attitudes like intention, goals, obligation [107, 79, 123] etc. They are also suited to formalize reasoning about actions and time [130] as well as in the specification of *interactions* among agents [19, 8, 85]. Moreover, they are used in theoretical computer science for program specification as well as verification [32, 93, 75]. The advantages of using modal logic over classical first-order logic in formalising cognitive notions like belief, intentions, etc. has been presented many times as in [131, 63, 85, 123].

The most basic and the one that is important from this thesis point of view is that modal languages could be seen as general tools for talking about relational structures[6]. A relational structure is simply a set together with a collection of relations on that. Modal languages are the simplest

[6]This view is sometimes called the Amsterdam-style perspective on modal logic and for a full exposition see [10].

languages in which relational structures can be described, constrained and reasoned about *locally*. The term *local* denotes that modal formulas are evaluated *inside* structures at a *particular* state.

> The function of the modal operators is to permit the information stored at other states to be scanned via an appropriate transition with the condition that there should be a current state from which these other states are accessible [10].

This *localised* perspective of modal languages on relational structures helps in giving highly intuitive mathematical meaning to a variety of modal operators as opposed to that of classical languages which have an *external* perspective on relational structures. Also from a combining logic point of view this internal perspective is important as satisfaction of formulas in the combined logic is done with respect to a particular state (world). Moreover, as it is possible to reduce modal logic to (fragments of) first order logic[7] it is possible to import and export various techniques and results available at both ends.

1.3.1 Fibring/Dovetailing

We adopt Gabbay's [40, 41, 42] fibring/dovetailing technique for combining BDI logics. The methodology allows the user to combine (*fibre*) the semantics of the two logics at hand and *weave* the two proof theories into a combined logic that preserves the basic properties of the components. In order to do so the two logics must be fully *presentable* through their syntax, semantics and proof-theory. If that is not the case the methodology is still usable. The strategy is then to extract the consequence relations from the two logics and finally fibre the relational semantics into the combined logic. The fibring methodology is sufficiently general to be applied to any combination.

1.4 Plan Of This Thesis

This thesis contains, in addition to the current one, five more chapters. In the next chapter we outline the syntax and semantics of modal logic and show their relationship with BDI logics. Two combining techniques called *fusion* and *fibring/dovetailing* are analysed and we argue that fibring is more suited for BDI-like logics than fusion.

[7] The study cf this sort of translation of any modal language into its corresponding classical language is known as *Correspondence Theory*.

In the third chapter, *Fibring semantics for* BDI *logics*, we first show that BDI logics without *interaction* axioms is a *dovetailed* system (i.e., dovetailing can be used as a semantic methodology to generate BDI logics). In order to accommodate interaction axioms we adopt Catach's multi-modal semantics which consists of a class of interaction axioms $G^{a,b,c,d}$ through which a wide class of modal systems including that of BDI could be generated. Further, we show that it is necessary to give *conditions* on the fibring function in order to combine BDI logics with interaction axioms. It is also shown that fibring as a combining technique could be extended to non-normal logics.

In the fourth chapter, *Extending* BDI *with composite actions*, we extend the BDI framework to accommodate *composite actions*. Three recent theories on *action* and *ability* are reviewed and it is argued that the notion of *ability* alone in BDI is not sufficient for reasoning about composite actions. It is shown that along with *ability*, notions like *opportunity* and *result* should be incorporated for the successful execution of such actions. Further we analyse the set of *commitment axioms* as given in [107] in the background of the new framework.

The fifth chapter advocates a non-monotonic logic for reasoning about intentions and as such deviates from the usual normal modal logic approach. Our view is based on Bratman's theory of *policy-based* intentions and we argue that policy-based intentions exhibit non-monotonic behavior which could be captured through a non-monotonic system like *defeasible logic*. The usual problems accompanying *logical omniscience* are avoided to a great extent by such an approach. We outline a proof-theory showing how intention-consistency could be maintained in BDI-like systems.

Finally, we conclude the thesis with a summary of the main contributions and a discussion on future research problems.

1.5 Bibliographic Note

Most of the research presented in the thesis has been previously published in some form. The main results of Chapter 3 on fibring semantics for BDI logics was presented as full paper [61] at the European Conference for logics in Artificial Intelligence (JELIA-02) in Italy. Chapter 4 is a longer version of the full paper presented at the fourteenth Australian Joint Conference for Artificial Intelligence (AI-01) in Adelaide (Australia). The results related to the defeasible nature of policy-based intentions reported in Chapter 5 appeared in a much shortened version in the proceedings of the Australian workshop on computational logic (AWCL-02) and was presented

at the workshop held in Canberra. It also appeared as a poster paper in the proceedings of the fifteenth Australian Joint Conference for Artificial Intelligence (AI-02) in Canberra (Australia).

Chapter 2

Intensional Logics and BDI

The limits of my language means the limits of my world; Logic pervades the world: the limits of the world are also its limits.

Ludwig Wittgenstein

As mentioned in the previous chapter, the underlying formal language in intentional systems like BDI is that of Modal Logic. In this chapter we give an overview of Modal Logic stating some of its important properties such as completeness and establish its relationship with BDI in a multi-modal setting. We then introduce two combining techniques called *fusion* and *fibring* and show why fibring is more suitable to combine BDI logics than fusion.

2.1 Modal Logic (The Basic Case)

Intentional notions like *belief, desires* and so forth are most commonly formulated by using a *modal logic* or some variants of it. In brief, the syntax of a modal language is that of classical languages (propositional or first-order) with an addition of non-truth functional operators. Many, however believe that anything that can be represented in non-classical languages such as modal logic, can be reduced to a problem in first-order logic. For instance, John McCarthy [92] questions the usefulness of modality in artificial intelligence (AI) and knowledge representation (KR). For McCarthy *[i]ntroducing new modalities should involve no more fuss than introducing a*

new predicate. But as Heinrich Wansing points out [125], using new predicates instead of modal operators deprives us of the appealing and insightful interpretation of modal formulas relative to possibly compound states with an interesting relational structure.

Secondly, from a combining logic point of view, though modal logics like **S4** and **S5** may be viewed as fragments of first-order logic, it is not the case that the reducing theory (first-order) is meant to replace the reduced theory (modal). It might be the case that such reductions may prove to be technically useful due to a transfer of known properties from the reducing theory to the reduced one. For instance, in description logics, which are restricted first-order formalisms, though the concept language $\mathcal{ALC}$ is only a syntactic variant of the smallest polymodal logic, it nevertheless has been suggested to further extend $\mathcal{ALC}$ by modal epistemic operators (see [86, 21]). Moreover, modal logic is no longer seen as just an extension of propositional logic, but also as a restriction of *first-order logic*. Other reasons against the use of classical logics for reasoning about intentional notions has been given in [131]. Thus modal logic is conceived as a much richer and more versatile research paradigm than interpreting normal modal operators as philosophically or otherwise interesting modalities. The new perspective on modal logic has enriched our theoretical understanding of *what modal logic is* by making use of new tools such as *translation and bi-simulation* and thereby showing which fragments of first-order logic they correspond to (see [10] for further details). It has encouraged modal logicians to think of themselves as *logic engineers*, whose task is to craft logics to fit particular applications, and this has lead to the development of new *extended modal logics*.

2.1.1 Syntax

The language of basic modal logic is that of propositional logic (**PL**) with two extra connectives, $\Box$ and $\Diamond$,[1]and denoted by $\mathcal{L}_{\mathbf{PML}}$ in this thesis. The *alphabet* of $\mathcal{L}_{\mathbf{PML}}$ consists of

- An enumerable set of *propositional variables*: $\Phi = \mathfrak{p}_0, \mathfrak{p}_1, \ldots, \mathfrak{q}_0, \mathfrak{q}_1 \ldots$;
- The *logical constants*: $\top$ (true) and $\bot$ (false);
- The *Boolean connectives*: $\wedge$ (and), $\vee$ (or), $\Rightarrow$ (implies), $\Leftrightarrow$ (if and only if) and $\neg$ (not);

[1]Though C.I. Lewis [82, 83] is considered as the creator of the first modal systems, it was Orlov and Godel who added it explicitly to classical propositional logic. (For details see [45]).

- The *Modal operators*: $\Box$ (*it is necessary*) and $\Diamond$ (*it is possible*)

Definition 1 *The language of basic modal logic is defined by the following Backus Naur Form (BNF):*

$$\mathcal{L}_{PML} ::= \bot \mid \top \mid \mathfrak{p}_i \mid \neg\varphi \mid (\varphi \wedge \psi) \mid (\varphi \vee \psi) \mid (\varphi \Rightarrow \psi)(\varphi \Leftrightarrow \psi) \mid \Box\varphi \mid \Diamond\varphi,$$

The propositional variables $(\mathfrak{p}_0, \mathfrak{p}_1, \ldots)$ and constants $(\bot, \top)$ denote atomic formulas ($\mathcal{L}_{\mathbf{PML}} - formula$) intended for representing compound propositions. Greek letters like φ, ψ, χ are reserved for formulas whereas Σ, Δ denotes sets of formulas. If φ and ψ are $\mathcal{L}_{\mathbf{PML}}$-*formulas* then so are $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$ etc. as with the standard tradition. Example formulas of basic modal logic are $(\mathfrak{p}_0 \wedge \Diamond(\mathfrak{p}_0 \Rightarrow \Box\neg\mathfrak{p}_3))$ and $\Box((\Diamond\mathfrak{p}_1 \wedge \neg\mathfrak{p}_3) \Rightarrow \Box\mathfrak{p}_0)$. Though many such formulas can be constructed using the basic modal language $\mathcal{L}_{\mathbf{PML}}$, not all of them constitute a modal *system*. The basic idea of constructing a modal system is to single out and describe those formulas that represent *true* propositions no matter what values are assigned to the variables. This set of formulas (usually referred to as *axioms*), together with a set of rules is denoted by Λ and is often called the logic of the corresponding system.

2.1.2 Semantics

As mentioned above, the formulas in a modal system should satisfy certain truth conditions. Hence we need a method for evaluating the truth or falsity of modal formulas. Though $\mathcal{L}_{\mathbf{PML}}$ is built on the language of classical propositional logic **PL**, it is not possible to adopt the classical semantics for $\mathcal{L}_{\mathbf{PML}}$ due to the operators $\Box$ and $\Diamond$. The reason is that in **PL** the only conditions to be accounted for is the truth/falsity of a variable in a *state* whereas in the case of $\Box\mathfrak{p}_0$ one has to think of truth or falsity of a variable with respect to *states* and certain *relations* between the states. Hence we need to interpret our modal language as a way of talking about *relational structures*. To this end we use *relational semantics* or *Kripke semantics* (*possible worlds semantics*) [78] to explicate the logical structure of our modal systems. This is done both at the level of *models* and at the level of *frames*. Both levels are important as they support the key notions of *satisfaction* and *validity*.

Definition 2 ***(Kripke frames and Kripke models)*** *A Kripke-frame is a pair* $\mathcal{F} = (\mathsf{W}, \mathsf{R})$ *where* W *is a set called the* ***set of worlds*** *and* R *is a binary relation called an* ***accessibility relation****. A frame is just a set of* points *and the relation between them. Whereas a Kripke model is a triple*

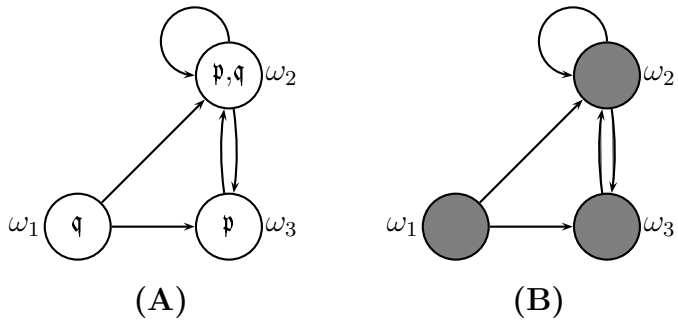


Figure 2.1: A Kripke model (A) and Kripke frame (B)

$\mathfrak{M} = (\mathcal{F}, \nu, \omega)$ *where* $\omega \in \mathsf{W}$ *and* ν *is a valuation on* $\mathcal{F}$*. A* ***valuation*** *is a function* $\nu : \Phi \to 2^{\mathsf{W}}$*. We say that the model* $\mathfrak{M}$ *is based on the frame* $\mathcal{F}$*, or that* $\mathcal{F}$ *is the underlying frame of* $\mathfrak{M}$*.*

In Figure 2.1 $\mathsf{W} = \{x_1, x_2, x_3\}$ represents the **set of possible worlds** and $x_1\mathsf{R}x_2$, $x_1\mathsf{R}x_3$, $x_2\mathsf{R}x_2$, $x_2\mathsf{R}x_3$ and $x_3\mathsf{R}x_2$ denote the **accessibility relations**. Elements of W are called *worlds, states* or *points* and $x_1\mathsf{R}x_2$ could be interpreted as x_2 is *accessible* from x_1 or x_1 *sees* x_2 or x_2 is a *successor* of x_1. A frame (as in Fig 2.1(**B**)) is more like a directed graph and has no information about the truth of atomic formulas at the various points. But in the case of models we have to evaluate the truth of propositions and hence need to introduce the notion of *satisfaction in a world* and this is defined inductively as follows:

1. $\mathfrak{M}, \omega \models \mathfrak{p}_0$ iff $\omega \in \nu(\mathfrak{p}_0)$
2. $\mathfrak{M}, \omega \models \neg\varphi$ iff $\mathfrak{M}, \omega \not\models \varphi$
3. $\mathfrak{M}, \omega \models \varphi \wedge \psi$ iff $\mathfrak{M}, \omega \models \varphi$ and $\mathfrak{M}, \omega \models \psi$
4. $\mathfrak{M}, \omega \models \Box\varphi$ iff for all $\omega' \in \mathsf{W}$ with $\omega\mathsf{R}\omega'$ implies $\mathfrak{M}, \omega' \models \varphi$

It follows from this definition that $\mathfrak{M}, \omega \models \Diamond\varphi$ iff for some $\omega' \in \mathsf{W}$ we have $\omega\mathsf{R}\omega'$ and $\mathfrak{M}, \omega' \models \varphi$. Satisfaction for the other logical connectives can be defined in the usual way.

Definition 3 ***(Satisfaction)*** *Let* $\mathfrak{M} = (\mathcal{F}, \nu, \omega)$ *be a model based on the frame* $\mathcal{F} = (\mathsf{W}, \mathsf{R})$*. A formula* φ *is said to be* ***globally true*** *in* $\mathfrak{M}$ *(in symbols:* $\mathfrak{M} \models \varphi$*) if it is satisfied at all states in* $\mathfrak{M}$ *(that is, if* $\mathfrak{M}, \omega \models \varphi$*, for all* $\omega \in \mathsf{W}$*, which means* $\nu(\varphi) = \mathsf{W}$*). Dually* φ *is* ***satisfied in*** $\mathfrak{M}$ *if* $\nu(\varphi)$ *is not empty.*

If we consider the Kripke model in figure 2.1(**A**) then we can find the following instances of the satisfaction condition:

- $x_1 \models \mathfrak{q}$, since $\mathfrak{q} \in \Sigma(x_1)$ where $\Sigma(x_1)$ is the set of all atomic formulas true at x_1;
- $x_1 \models \Diamond\mathfrak{q}$ for there is a world related to x_1 (i.e. x_2) which satisfies $\mathfrak{q}$;
- $x_1 \not\models \Box\mathfrak{q}$ because $x_1 \models \Box\mathfrak{q}$ means that all worlds related to x_1 (ie. x_2 and x_3) satisfy $\mathfrak{q}$; but x_3 does not;
- $x_2 \models \Box\mathfrak{p} \Rightarrow \mathfrak{p}$ and $x_3 \models \Box\mathfrak{p} \Rightarrow \mathfrak{p}$. $x_1 \not\models \Box\mathfrak{p} \Rightarrow \mathfrak{p}$ because though it satisfies $\Box\mathfrak{p}$ it does not satisfy $\mathfrak{p}$;

Definition 4 *(**Validity**) We say that a formula φ is valid in a frame $\mathcal{F}$ (or $\mathcal{F}$ validates φ) and write $\mathcal{F} \models \varphi$, if for every valuation ν, $\nu(\varphi) = \mathsf{W}$, i.e., if φ is true in all models based on $\mathcal{F}$. And φ is* satisfiable in $\mathcal{F}$*, if it is satisfiable in some model based on $\mathcal{F}$. A formula φ is valid on a class of frames $\mathfrak{F}$ (notation: $\mathfrak{F} \models \varphi$) if it is valid on every frame $\mathcal{F}$ in $\mathfrak{F}$. For a set Γ of $\mathcal{L}_{\mathbf{PML}}$-formulas, we say that $\mathcal{F}$ is a* frame for Γ *if all formulas in Γ are valid in $\mathcal{F}$ and write $\mathcal{F} \models \Gamma$. A formula φ is* Γ-satisfiable *if it is satisfiable in a frame for Γ. The set of all formulas valid in a class of frames $\mathfrak{F}$ is called the* logic *of $\mathfrak{F}$.*

The above definition establishes a connection between logics and frames which is helpful in determining the semantical characterization of some modal logics. For instance if we consider $\mathfrak{F}$ to be an arbitrary class of frames then,

$$\Delta_{\mathfrak{F}} = \{\varphi \in \mathcal{L}_{\mathbf{PML}} \mid \forall \mathcal{F} \in \mathfrak{F},\ \mathcal{F} \models \varphi\} \tag{2.1}$$

is a modal logic called the *logic of* $\mathfrak{F}$. We talk more about this relationship in the next section when we discuss soundness and completeness of modal logics. So far we have seen many different variations on the concept of *truth* or *satisfaction* from a semantic (*possible worlds semantics*) point of view, viz.

- truth of a formula at a world of a model,
- truth of a formula in a model,
- truth of a formula in a frame,
- truth of a formula in a class of frames.

But there is also a *syntactic* way of defining a logic with the help of *inference systems* like *Hilbert-style calculi*. To define such a system one has to indicate which formulas are regarded as *axioms* of the system and then specify its *inference rules*. A *derivation of a formula* φ in such systems is a finite sequence of formulas ending with φ, such that, each formula in the sequence is either an axiom or obtained from earlier formulas in the sequence by applying one of the inference rules. A logic defined in this way is the smallest set of formulas which contains the axioms and is closed under the inference rules. For instance, it is well known that classical propositional logic **PL** can be defined using a particular Hilbert-style calculus.

Usually it is desirable that the semantical and syntactical definitions complement each other as the former explains the (intended) meaning of the logical constants and connectives while the latter provides us with the reasoning machinery. Hence, if the reasoning principles of **PL** are accepted, it is possible to come up with a modal calculi by adding to the Hilbert-style calculus of **PL** those axioms and inference rules describing the *additional* modal operators. The usual question is as to whether there is a syntactic mechanism capable of generating $\Lambda_{\mathfrak{F}}$, the formulas valid on $\mathfrak{F}$, for a certain class of frames. Such a mechanism is best explained in the concept of **Normal modal logic (NML)**. A Normal modal logic Λ is a subset of formulas of basic modal logic $\mathcal{L}_{\mathbf{PML}}$, as specified in Definition 1, and can be deduced from a set of axioms and inference rules. The set of axioms include the following:

1. All instances of the propositional calculus
2. (Axiom **K**) $\Box(\mathfrak{p}_0 \Rightarrow \mathfrak{p}_1) \Rightarrow (\Box\mathfrak{p}_0 \Rightarrow \Box\mathfrak{p}_1)$
3. $\Diamond\mathfrak{p}_0 \Leftrightarrow \neg\Box\neg\mathfrak{p}_0$

The set of inference rules are as follows:

- Modus Ponens (MP): if φ and $\varphi \Rightarrow \psi$ are theorems so is ψ
- Necessitation (NEC): if φ is a theorem, so is $\Box\varphi$
- Uniform Substitution (Subst): given a formula $\varphi(\mathfrak{p}_1, \ldots, \mathfrak{p}_n)$, derive the formula $\varphi\{\psi_1/\mathfrak{p}_1, \ldots, \psi_n/\mathfrak{p}_n\}$ which is obtained by uniformly substituting formulas $\psi_1 \ldots \psi_n$ instead of the variables $\mathfrak{p}_1, \ldots, \mathfrak{p}_n$ in φ respectively.

Definition 5 *(**Normal Modal Logics**) A normal modal logic* Λ *is a set of modal formulas that contains all propositional tautologies,* $\Box(\mathfrak{p} \Rightarrow \mathfrak{q}) \Rightarrow$

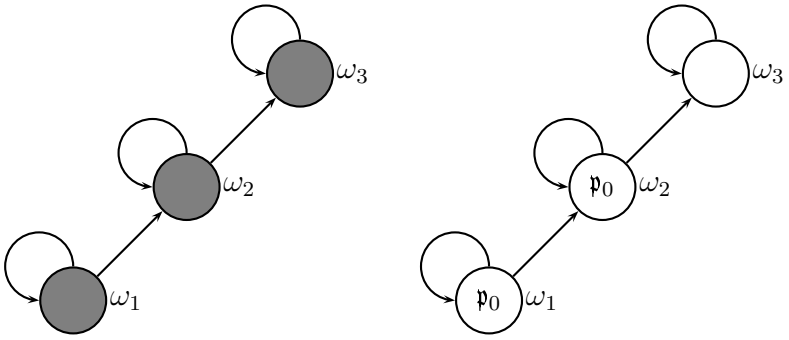
$(\Box\mathfrak{p} \Rightarrow \Box\mathfrak{q})$, $\Diamond\mathfrak{p} \Leftrightarrow \neg\Box\neg\mathfrak{p}$, *and that is closed under necessitation (that is, if* $\vdash_\Lambda \varphi$ *then* $\vdash_\Lambda \Box\varphi$ *where ,* $\vdash_\Lambda \varphi$ *means* φ *is a theorem of* Λ*), modus ponens and uniform substitution.*

The smallest normal modal logic is called **K** as it just contains propositional logic and all instances of the formula schema **K**, together with other formulas from applying inference rules like modus ponens, necessitation and uniform substitution. Every other normal modal logic Λ can be obtained by extending this system with a set Σ of *axioms*, denoted as $\Lambda = \mathbf{K} \oplus \Sigma$. If Σ is finite then Λ is called *finitely axiomatisable* and for a given normal logic Λ and a set Σ of $\mathcal{L}_{\mathbf{PML}}$ formulas $\Lambda \oplus \Sigma$ is the smallest normal logic containing $\Lambda \cup \Sigma$ and we denote $\Lambda \oplus \varphi$ if $\Sigma = \{\varphi\}$. Table 2.1 shows a list of logics thus obtained.

Name of the Logic	Logic	Axiom Name
T	$\mathbf{K} \oplus \Box\mathfrak{p}_0 \Rightarrow \mathfrak{p}_0$	**T**
KD	$\mathbf{K} \oplus \Box\mathfrak{p}_0 \Rightarrow \Diamond\mathfrak{p}_0$	**D**
K4	$\mathbf{K} \oplus \Box\mathfrak{p}_0 \Rightarrow \Box\Box\mathfrak{p}_0$	**4**
S4	$\mathbf{K} \oplus \Box\mathfrak{p}_0 \Rightarrow \Box\Box\mathfrak{p}_0 \oplus \Box\mathfrak{p}_0 \Rightarrow \mathfrak{p}_0$	**4, T**
S5	$\mathbf{S4} \oplus \Diamond\mathfrak{p}_0 \Rightarrow \Box\Diamond\mathfrak{p}_0$ $\mathbf{S4} \oplus \mathfrak{p}_0 \Rightarrow \Box\Diamond\mathfrak{p}_0$	**5** **B**
S4.2	$\mathbf{S4} \oplus \Diamond\Box\mathfrak{p}_0 \Rightarrow \Box\Diamond\mathfrak{p}_0$	**G**
KD45	$\mathbf{K4} \oplus \Box\mathfrak{p}_0 \Rightarrow \Diamond\mathfrak{p}_0 \oplus \Diamond\mathfrak{p}_0 \Rightarrow \Box\Diamond\mathfrak{p}_0$	**S5**
Triv	$\mathbf{K} \oplus \mathfrak{p} \Rightarrow \Box\mathfrak{p}$	**Triv**

Table 2.1: Some important logics

Though we defined that there is a correspondence between frames and some of the modal logics (2.1) we were not specific in pointing out what kind of frames correspond to the different modal logics. Given below is an example that shows, for a given frame $\mathcal{F} = (\mathsf{W}, \mathsf{R})$, the properties R should have in order to satisfy a given formula. For instance, consider the frame $\mathcal{F}$ (1A) given in Figure 2.2. $\mathcal{F}$ satisfies the formula $\Box\mathfrak{p}_0 \Rightarrow \mathfrak{p}_0$. Though a frame has no information about what atomic formulas are true at various points, it could be said that a frame as a whole satisfies a formula. In order to show the satisfaction of the above formula we have to show that for any world ω if $\omega \models \Box\mathfrak{p}_0$ then $\omega \models \mathfrak{p}_0$. Since $\omega\mathsf{R}\omega$, as can be seen from the diagram, it follows from the definition of $\Box$ that $\omega \models \mathfrak{p}_0$ (since $\omega \models \Box\mathfrak{p}_0$ iff for each $\omega' \in \mathsf{W}$ we have $\omega\ \mathsf{R}\ \omega'$ implies $\omega' \models \mathfrak{p}_0$).



(1A) **(1B)**

Figure 2.2: A Kripke frame satisfying the formula $\Box\mathfrak{p}_0 \Rightarrow \mathfrak{p}_0$ (1A) and the associated Kripke Model (1B)

Therefore $\mathcal{F}$ satisfies the schema $\Box\mathfrak{p}_0 \Rightarrow \mathfrak{p}_0$. It is also equivalent to say that since the relation R is *reflexive*[2] the frame satisfies the schema. But $\mathcal{F}$ does not satisfy the formula $\Box\mathfrak{p}_0 \Rightarrow \Box\Box\mathfrak{p}_0$ because if we consider (1B) then $\omega_1 \models \mathfrak{p}_0$, but $\omega_1 \not\models \Box\Box\mathfrak{p}_0$. In other words we can say that R is not *transitive*[3].

To sum up, we can prove that

1. If R has the property then the frame satisfies the formula scheme.
2. If the frame satisfies the schema then it satisfies an instance of it and
3. If the frame satisfies the formula, then R has the property.

Table 2.2 shows the property of R corresponding to a collection of formulas.

Theorem 1 *A frame* $\mathcal{F} = (\mathsf{W}, \mathsf{R})$ *satisfies an axiom scheme in table 2.2 iff* R *has the corresponding property in that table.*

Theorem 2 *A frame* $\mathcal{F} = (\mathsf{W}, \mathsf{R})$ *satisfies an axiom scheme in table 2.2 iff* R *has the corresponding property in that table.*

The theorem is important for it helps one in choosing the specific axiom schema relevant for his/her application domain by looking at the properties of R. A *correspondence* can be drawn between formulae that are *valid* on specific classes of frames and *structural* properties of the frames themselves[4]. This means that under certain circumstances syntax and semantics

[2] R is reflexive if $\forall\ \omega \in \mathsf{W}\ \omega\mathsf{R}\omega$.

[3] R is transitive if $\forall\omega, \omega', \omega'' \in \mathsf{W}(\omega\mathsf{R}\omega' \wedge \omega'\mathsf{R}\omega'' \Rightarrow \omega\mathsf{R}\omega'')$.

[4] This is the subject matter of *Correspondence theory* [122].

Frame property	Property Name	Axiom
$\forall\omega \in \mathsf{W}\ \omega\mathsf{R}\omega$	Reflexive	$\Box\varphi \Rightarrow \varphi$
$\forall\omega \in \mathsf{W}\ \exists\omega'\ \omega\mathsf{R}\omega'$	Serial	$\Box\varphi \Rightarrow \Diamond\varphi$
$\forall\omega,\omega',\omega'' \in \mathsf{W}\ (\omega\mathsf{R}\omega' \wedge \omega'\mathsf{R}\omega'' \Rightarrow \omega\mathsf{R}\omega'')$	Transitive	$\Box\varphi \Rightarrow \Box\Box\varphi$
$\forall\omega,\omega' \in \mathsf{W}\ (\omega\mathsf{R}\omega' \Rightarrow \omega'\mathsf{R}\omega)$	Symmetric	$\varphi \Rightarrow \Box\Diamond\varphi$
$\forall\omega,\omega',\omega'' \in \mathsf{W}\ (\omega\mathsf{R}\omega' \wedge \omega\mathsf{R}\omega'' \Rightarrow \omega'\mathsf{R}\omega'')$	Euclidean	$\Diamond\varphi \Rightarrow \Box\Diamond\varphi$
$\forall\omega\ \omega\mathsf{R}\omega \wedge \forall\omega'(\omega\mathsf{R}\omega' \Rightarrow \omega = \omega')$	Reflexive dead-end	$\varphi \Leftrightarrow \Box\varphi$
$\forall\omega,\omega',\omega'' \in \mathsf{W}\ (\omega\mathsf{R}\omega' \wedge \omega\mathsf{R}\omega'' \Rightarrow \omega' = \omega'')$	Functional	$\Diamond\varphi \Rightarrow \Box\varphi$
$\forall\omega,\omega',\omega'' \in \mathsf{W}\ (\omega\mathsf{R}\omega' \wedge \omega\mathsf{R}\omega'' \Rightarrow$ $\exists\omega'''(\omega'\mathsf{R}\omega''' \wedge \omega''\mathsf{R}\omega'''))$	Convergent	$\Diamond\Box\varphi \Rightarrow \Box\Diamond\varphi$

Table 2.2: The property of R corresponding to some formulas

in modal logic are interchangeable and leads us to the fundamental concepts of *soundness* and *completeness* which links the syntactic and semantic perspectives. It should be noted that working out the formulae corresponding to some properties can be done in an automatic fashion by means of a result by Sahlqvist [110]. The algorithm calculates the frame condition associated to particular formulas. Moreover, it is possible to guess frame correspondence once some expereince is gained.

Before examining the notions of soundness and completeness we introduce another important concept that needs mention in relation to frames: *morphism* (*structure preserving maps*), which helps in establishing invariance results. The usual practice is to identify the kind of morphisms suitable for a particular class of frames/models and check whether they give rise to invariance results (i.e., to check whether a particular class of frames/models can be made to do the work of another class of frames/models). We consider two kinds of morphism here and for further reference see [10].

Definition 6 *Two frames $\mathcal{F} = (\mathsf{W}, \mathsf{R}), \mathcal{F}' = (\mathsf{W}', \mathsf{R}')$ in a class $\mathfrak{F}$ of frames are isomorphic ($\mathcal{F} \cong_{\mathfrak{F}} \mathcal{F}'$) if and only if*

- *There exists a bijection $b : \mathsf{W} \mapsto \mathsf{W}'$;*
- *For all $\omega, \omega' \in \mathsf{W}, \omega\mathsf{R}\omega'$ if and only if $b(\omega)\mathsf{R}'b(\omega')$.*

The models, $(\mathsf{W}, \mathsf{R}, \nu)$ and $(\mathsf{W}', \mathsf{R}', \nu')$ are said to be isomorphic if and only if (W, R) and $(\mathsf{W}', \mathsf{R}')$ are isomorphic frames, and for any variable $\mathfrak{p}_0$ and any $\omega \in \mathsf{W}, \nu(\mathfrak{p}_0, \omega) = \nu'(\mathfrak{p}_0, b(\omega))$.

Lemma 1 *If $\mathcal{F}$ and $\mathcal{F}'$ are isomorphic frames then for all φ we have $\mathcal{F} \models \varphi$ if and only if $\mathcal{F}' \models \varphi$.*

Frame isomorphism, as defined above, is structure preserving in a strong sense. A slightly weaker property called *pseudo-epimorphism* or *p-morphism* is given by Segerberg and it could be defined both at the level of frames and at the level of Kripke models.

Definition 7 ***(p-morphisms)*** *Let* $\mathfrak{M} = (\mathsf{W}, \mathsf{R}, \nu)$ *and* $\mathfrak{M}' = (\mathsf{W}', \mathsf{R}', \nu')$ *be models, and* $p : \mathsf{W} \mapsto \mathsf{W}'$ *a function satisfying*

1. *p is a function from* W *onto* W' *(p is surjective);*
2. *for all* $\omega, \omega' \in \mathsf{W}$ $\omega \mathsf{R} \omega'$ *implies* $p(\omega) \mathsf{R}' p(\omega')$*;*
3. *For any* $\omega \in \mathsf{W}$ *and* $\omega'' \in \mathsf{W}'$ $p(\omega) \mathsf{R}' \omega''$ *implies* $\exists \omega' \in \mathsf{W}(\omega \mathsf{R} \omega' \wedge p(\omega') = \omega''$*.*
4. *for every proposition* $\mathfrak{p}_0$ *and every world* $\omega \in \mathsf{W}$ $\omega \in \nu(\mathfrak{p}_0)$ *if and only if* $p(\omega) \in \nu'(\mathfrak{p}_0)$

then p is called a *model* p-morphism from $\mathfrak{M}$ to $\mathfrak{M}'$. A function satisfying the first three conditions is a *frame* p-morphism from frame (W, R) to frame $(\mathsf{W}', \mathsf{R}')$. If there is a p-morphism from $\mathcal{F}$ to $\mathcal{F}'$, $\mathcal{F}'$ is also said to be a p-morphic image of $\mathcal{F}$.

Lemma 2 *[54] If p is a model p-morphism from* $\mathfrak{M}$ *to* $\mathfrak{M}'$ *then for all worlds* ω *of* $\mathfrak{M}$ *and formulae* $\varphi \in \mathcal{L}_{\boldsymbol{PML}}$*, we have* $\mathfrak{M}, \omega \models \varphi$ *if and only if* $\mathfrak{M}', p(\omega) \models \varphi$*. A formula* φ *is valid in* $\mathfrak{M}$ *if and only if* φ *is valid in* $\mathfrak{M}'$*. If p is a frame p-morphism from* $\mathcal{F}$ *to* $\mathcal{F}'$ *then for all* $\varphi \in \mathcal{L}_{\boldsymbol{PML}}$*, if* $\mathcal{F} \models \varphi$ *then* $\mathcal{F}' \models \varphi$*.*

The above lemma shows that p-morphisms preserve satisfaction and validity for the language $\mathcal{L}_{\mathbf{PML}}$.

2.1.3 Soundness, Completeness and Canonical Models

So far we have come across the syntactic and semantic specification of modal logics, in particular *normal modal logics*, where we have seen that NML's are sets of formulae satisfying certain closure conditions. We have seen how they can be specified semantically through *relational semantics* and syntactically through a *Hilbert-style* inference system. It is often asked whether a syntactically specified logic can be characterised on a semantic basis and vice versa; this has led to the development of several formal techniques such as completeness proofs via canonical models [28, 89, 80],

decidability via filtrations [80] etc. Recently completeness and decidability transfer via combining techniques like *fusion* and *fibring* [77, 42] have been proposed.

In this thesis we make use of the *canonical model* construction to show the individual properties of BDI logics and *fibring* as the semantic methodology to generate such combined logics. Before going into the details we have to fix few notations which would be used frequently in this thesis.

In the previous section we defined Σ to be a set of axioms, when added to the **K** axiom can give rise to different normal logics. It is often the case that Σ could also be referred to as a set of *formulae* that generates a logic, which is the usual way of syntactically specifying a normal logic. (Since Hibert-style terminology is used more often, Σ is referred to as *axioms*). The consistency conditions for such sets of formulae is defined as follows:

Definition 8 ***(Maximal consistency)*** *Let $\Sigma \subseteq \mathcal{L}_{PML}$. We say that Σ is Λ-inconsistent if there are $\psi_1, \ldots, \psi_n \in \Sigma$ such that $\Lambda \vdash \neg(\psi_1 \wedge \ldots \wedge \psi_n)$. If that is not the case, Σ is said Λ-consistent. A formula φ is Λ-consistent if $\{\varphi\}$ is. An infinite set of formulae Σ is Λ-consistent if any finite subset of Σ is Λ-consistent. A set Σ of formulas is* ***maximal*** *if for every formula φ, either $\varphi \in \Sigma$ or $\neg\, \varphi \in \Sigma$. A set Σ is Λ-**maximal consistent** if it is both Λ-consistent and maximal (denoted by $Max_\Lambda\Sigma$ with the reading Σ is Λ-maximal).*

Lemma 3 *Let $\Sigma \subseteq \mathcal{L}_{PML}$ be a maximal Λ-consistent set. Then the following properties hold:*

1. *For any $\varphi, \psi \in \mathcal{L}_{PML} : \varphi \vee \psi \in \Sigma$ iff either $\varphi \in \Sigma$, or $\psi \in \Sigma$.*
2. *For any $\varphi, \psi \in \mathcal{L}_{PML} : \varphi \wedge \psi \in \Sigma$ iff both $\varphi \in \Sigma$, and $\psi \in \Sigma$.*
3. *Σ is closed under modus ponens.*
4. *$\Lambda \subseteq \Sigma$ (i.e., for any $\varphi \in \mathcal{L}_{PML}, \varphi \in \Lambda$).*

Lemma 4 ***(Lindenbaum's Lemma)*** *If Σ is a Λ-consistent set of formulas ($Con_\Lambda\Sigma$) then there is a Λ-maximal consistent set Δ ($Max_\Lambda\Delta$) such that $\Sigma \subseteq \Delta$.*

Lemma 5 *Let Λ be a logic and Σ a maximal Λ-consistent set such that $\Diamond\varphi \in \Sigma$. Then there exists a maximal Λ-consistent set Δ such that $\{\psi \mid \Box\psi \in \Sigma\} \cup \{\varphi\} \subseteq \Delta$.*

Proof. We first prove that $\Delta^{-} = \{\psi \mid \Box\psi \in \Sigma\} \cup \{\varphi\}$ is Λ-consistent. By contradiction, suppose that is not the case. Then there must exist $\psi_1, \ldots, \psi_n \in \mathcal{L}_{\mathbf{PML}}$ such that $\Lambda \vdash \neg(\psi_1 \wedge \ldots \wedge \psi_n \wedge \varphi)$. So by propositional calculus $\Lambda \vdash (\psi_1 \wedge \ldots \wedge \psi_n) \Rightarrow \neg\varphi$. But since Λ is *normal*, $\Lambda \vdash \Box(\psi_1 \wedge \ldots \wedge \psi_n) \Rightarrow \Box\neg\varphi$. By the distributivity property, $\Box$ distributes over $\wedge$ and so $\Lambda \vdash (\Box\psi_1 \wedge \ldots \wedge \Box\psi_n) \Rightarrow \Box\neg\varphi$. But by hypothesis $\Box\psi_1 \in \Sigma, \ldots, \Box\psi_n \in \Sigma$ and therefore $\Box\neg\varphi \in \Sigma$ (because if $\psi_1 \in \Sigma$ and $\Lambda \vdash \psi_1 \Rightarrow \psi_2$, then $\psi_2 \in \Sigma$). From this it follows that $\neg\Diamond\varphi \in \Sigma$ and thereby $\bot \in \Sigma$ which is absurd because Σ is assumed consistent.
Now consider Δ^{-}. Then by **Lindenbaum's Lemma**, there exists a maximal Λ-consistent extension and this is what was left to be shown. □

The concept of Λ-maximal consistent set of formulas ($Max_{\Lambda}\Sigma$) is used to prove completeness via *canonical models*, which are special models, whose worlds are all Λ-maximal consistent set of formulas. This means that if φ is true in some canonical model for Λ, then φ belongs to a Λ-maximal consistent set. We deal more with canonical models when we prove completeness for a logic.

Definition 9 *(**Soundness**) A normal modal logic Λ is **sound** with respect to a class $\mathfrak{F}$ of frames (or $\mathfrak{F}$-sound), if $\Lambda \subseteq \Delta_{\mathfrak{F}}$ (i.e., if $\mathcal{F} \models \varphi$ for all $\varphi \in \Lambda$ and all $\mathcal{F} \in \mathfrak{F}$). If Λ is sound with respect to $\mathfrak{F}$ then it is said that $\mathfrak{F}$ is a class of frames (or models, or general frames) for Λ*

Definition 10 *(**Completeness**) A normal modal logic Λ is **complete** with respect to a class $\mathfrak{F}$ of frames ($\mathfrak{F}$-complete), if $\Delta_{\mathfrak{F}} \subseteq \Lambda$ (i.e., if for any formula $\varphi \in \Lambda$ we have that $\mathfrak{F} \models \varphi$ implies $\Lambda \vdash \varphi$). We say that Λ is **determined** (or **characterised b**) by $\mathfrak{F}$ if Λ is both $\mathfrak{F}$-sound and $\mathfrak{F}$-complete, i.e., $\Lambda = \Delta_{\mathfrak{F}}$.*

While defining *maximal consistent sets* we mentioned that they are helpful in defining completeness proofs via canonical models. Using the *canonical model* construction is one of the easiest ways to prove completeness. It should be noted that completeness is defined with respect to a class of frames, not with respect to a class of models.

Definition 11 *(**Canonical model**) The **canonical model** $\mathfrak{M}^{\Lambda}_{\zeta}$ for a normal modal logic Λ is the triple $(\mathsf{W}^{\Lambda}_{\zeta}, \mathsf{R}^{\Lambda}_{\zeta}, \nu^{\Lambda}_{\zeta})$ where:*

- $\mathsf{W}^{\Lambda}_{\zeta}$ *is the set of all $Max_{\Lambda}\Sigma$*
- $\mathsf{R}^{\Lambda}_{\zeta}$ *the **canonical relation**, is a binary relation on W^2 such that $\omega\mathsf{R}^{\Lambda}_{\zeta}\omega'$ if and only if for all φ: $\Box\varphi \in \omega$ implies $\varphi \in \omega'$*

- ν_ζ^Λ *is the* ***canonical valuation*** *defined by* $\nu^\Lambda(\mathfrak{p}_0) = \{\omega \in \mathsf{W}^\Lambda \mid \mathfrak{p}_0 \in \omega\}$.

The pair $\mathcal{F}_\zeta^\Lambda = (\mathsf{W}^\Lambda, \mathsf{R}^\Lambda)$ *is called the* ***canonical frame*** *for* Λ. *It is to be noted that the definition of* R_ζ^Λ *is also equivalent to:*

$$\omega \mathsf{R}_\zeta^\Lambda \omega' \textit{ iff for all } \varphi : \varphi \in \omega' \textit{ implies } \Diamond\varphi \in \omega.$$

The basic idea behind canonicity is that, given a class of frames $\mathfrak{F}$, and a Hilbert system Λ, defined on a modal language $\mathcal{L}_{\mathbf{PML}}$, can we find $\mathfrak{F} \models \varphi \Rightarrow \Lambda \vdash \varphi$? For this to happen we have to check if it is possible to built a model $\mathfrak{M}_\zeta^\Lambda = (\mathcal{F}_\zeta^\Lambda, \nu)$, such that for any $\varphi \in \mathcal{L}_{\mathbf{PML}}$ we have that $\mathcal{F}_\zeta^\Lambda \models \varphi$ if and only if $\Lambda \vdash \varphi$. Here the worlds are particular *sets of formulas*, and prove $\mathfrak{M}_\zeta^\Lambda, \omega \models \varphi$ iff $\varphi \in \omega$.

The canonical model will have the property that $\mathfrak{M}_\zeta^\Lambda \models \varphi$ if and only if $\Lambda \vdash \varphi$. If we can establish that $\mathcal{F}_\zeta^\Lambda \in \mathfrak{F}$ then we can reason $\mathfrak{F} \models \varphi$ implies $\mathcal{F}_\zeta^\Lambda \models \varphi$. So $\mathfrak{M}_\zeta^\Lambda \models \varphi$ and therefore $\Lambda \vdash \varphi$. In this way we can prove completeness for Λ with respect to $\mathfrak{F}$. An important point to note is that we cannot have *any* formulas in the worlds, as we risk having $\bot \in \omega$, which would entail $\Lambda \vdash \bot$. Hence, as mentioned earlier the set of formulas should be $Max_\Lambda \Sigma$ (maximal consistent sets).

Lemma 6 *(**Truth Lemma**) For any normal modal logic* Λ *and any formula* φ, $\mathfrak{M}_\zeta^\Lambda, \omega \models \varphi$ *if and only if* $\varphi \in \omega$.

Proof. The proof is by structural induction on φ (the atomic case is satisfied by definition, and the propositional connectives can be verified as usual). The inductive step for the modalities is as follows:

1. Suppose $\Box\varphi \in \omega$, it remains to prove $\mathfrak{M}_\zeta^\Lambda, \omega \models \Box\varphi$, i.e., that for any ω' such that $\omega \mathsf{R}_\zeta^\Lambda \omega'$ we have $\mathfrak{M}_\zeta^\Lambda, \omega' \models \varphi$. But by induction hypothesis this is equivalent to proving $\varphi \in \omega'$. In turn, this is satisfied because of the definition of R_ζ^Λ and of the assumption $\Box\varphi \in \omega$. So, $\mathfrak{M}_\zeta^\Lambda, \omega \models \Box\varphi$.

2. Suppose $\Box\varphi \notin \omega$. So, by maximality, we have that $\Diamond\neg\varphi \in \omega$. So by Lemma 2 there exists a ω' such that $\{\psi \mid \Box\psi \in \omega\} \cup \neg\varphi \subseteq \omega'$; but then $\omega \mathsf{R}_\zeta^\Lambda \omega'$ and by induction hypothesis $\mathfrak{M}_\zeta^\Lambda, \omega' \models \neg\varphi$. So we have that $\mathfrak{M}_\zeta^\Lambda, \omega \models \neg\Box\varphi$

□

The lemma above leads to the following result

Corollary 1 *For any logic* Λ*:* $\mathfrak{M}^{\Lambda}_{\zeta} \models \varphi$ *if and only if* $\Lambda \models \varphi$*.*

Proof. Suppose $\Lambda \models \varphi$. Then φ is in every maximal Λ-consistent set; so $\varphi \in \omega$ and $\omega \in \mathsf{W}^{\Lambda}_{\zeta}$. So by Lemma 3 $\mathfrak{M}^{\Lambda}_{\zeta}, \omega \models \varphi$ for any $\omega \in \mathsf{W}^{\Lambda}_{\zeta}$, i.e., $\mathfrak{M}^{\Lambda}_{\zeta} \models \varphi$. Suppose $\Lambda \nvdash \varphi$. Then $\neg\varphi$ is Λ-consistent. So there exists a world ω such that $\neg\varphi \in \omega$. But then $\mathfrak{M}^{\Lambda}_{\zeta}, \omega \models \neg\varphi$, so $\mathfrak{M}^{\Lambda}_{\zeta} \not\models \varphi$. □

The corollary above shows that all and only the theorems of Λ are valid on a particular semantic structure: the canonical model $\mathfrak{M}^{\Lambda}_{\zeta}$. But since $\mathfrak{M}^{\Lambda}_{\zeta}$ is an infinite model and its relational properties are not known apriori the need arises to move toward results that establish one-to-one correspondence between theorems of a system and formulas valid in a class of frames. The canonical model helps in doing so.

Theorem 3 *(**Completeness via canonical model**) Let* Λ *be a logic and let* $\mathfrak{F}$ *be a class of frames. If the frame* $\mathcal{F}^{\Lambda}_{\zeta}$ *underlying the canonical model* $\mathfrak{M}^{\Lambda}_{\zeta}$ *for* Λ *is in the class* $\mathfrak{F}$ *then the logic* Λ *is complete with respect to* $\mathfrak{F}$

Theorem 4 ***K*** *is complete with respect to all frames.*

Theorem 5 *(**Completeness of basic logics**) For the properties and logics shown in Table 2.3 we have that:* Λ *is sound and complete with respect to the class of frames that have the corresponding property.*

Frame property	Logic Λ_i
Reflexive	**T**
Serial	**KD**
Transitive	**4**
Reflexive and Transitive (quasi-ordered)	**S4**
Reflexive, Transitive and Convergent	**S4.2**
Reflexive, Symmetric and Transitive	**S5**
Serial and Transitive	**KD4**
Serial, Transitive and Euclidean	**KD45**
Functional	**Alt**
Reflexive dead-end	**Triv**

Table 2.3: Completeness table for some of the basic Logics

Some of the definitions are as follows; A transitive and reflexive relation on W is called a *quasi-order* on W and a symmetric quasi-order is called an *equivalence relation.* $\mathcal{F}$ is a *quasi-ordered frame* if R is quasi-order on

W. A transitive, reflexive and antisymmetric R is called a *partial order*. An irreflexive and transitive relation is known as *strict partial order*. It could be observed that by imposing various restrictions on the accessibility relation R we can obtain different *systems* of propositional modal logic as shown in Table 2.3.

2.1.4 The Schema $G^{k,l,m,n}$

The work by Lemmon and Scott [80] on canonical models and completeness-via-canonicity arguments, provided some important results, of which, one was a general canonicity result for axioms of the form $\Diamond^k\Box^l\varphi \Rightarrow \Box^m\Diamond^n\varphi$. In a later work Catach [19] proposed a set of interaction axioms based on such a general scheme and proved completeness for the normal multi-modal systems generated by the interaction axioms. The schema $G^{k,l,m,n}$ is a generalisation of an axiom scheme $\Diamond\Box\varphi \Rightarrow \Box\Diamond\varphi$ which corresponds to the frame property

$$\forall\omega,\omega',\omega'' \in \mathsf{W}(\omega\mathsf{R}\omega' \wedge \omega\mathsf{R}\omega'' \Rightarrow \exists\omega'''(\omega'\mathsf{R}\omega''' \wedge \omega''\mathsf{R}\omega''')) \qquad (2.2)$$

When a relation R has this property it is called *incestual* (or confluent), since it means that offspring ω' and ω'' of a common parent ω have themselves an offspring ω''' in common.

Definition 12 *Let ω and ω' be worlds in a standard model $\mathfrak{M} = (\mathsf{W}, \mathsf{R}, \nu)$*

- $\omega\mathsf{R}^0\omega'$ *iff* $\omega = \omega'$
- *For $n > 0$, $\omega\mathsf{R}^n\omega'$ iff for some ω'' in $\mathfrak{M}$, $\omega\mathsf{R}\omega''$ and $\omega''\mathsf{R}^{n-1}\omega'$*

where the relation R^n is called the n-th *relative product* of R. From the above definition we can come to the conclusion that R^0 is the relation of identity, R^1 is the relation R itself and for $\omega\ \mathsf{R}^3\ \omega'$ iff $\exists\ \omega''$, ω''' such that $\omega\mathsf{R}\omega''$, $\omega''\mathsf{R}\omega'''$ and $\omega'''\mathsf{R}\omega'$. The truth conditions for sentences of the form $\Box^n A$ and $\Diamond^n A$ could be stated using R^n as $\Box A$ is true at a world exactly when A is true at all worlds lying n steps from it. Hence $\Box^n A$ holds at a world just in case A holds at all worlds n times removed from it. This condition could be generalised to say that R is $k, l, m, n - incestual$ if and only if for every ω, ω', and ω'' in $\mathfrak{M}$,

$$\text{if } \omega\mathsf{R}^k\omega' \text{ and } \omega\mathsf{R}^m\omega'', \text{ then } \exists\omega''' \text{ in } \mathfrak{M},\ \omega'\mathsf{R}^l\omega''' \text{ and } \omega''\mathsf{R}^n\omega'''$$

To be more general, when $\omega\mathsf{R}\omega'$ in a standard model $(\mathsf{W}, \mathsf{R}, \nu)$ we say that ω' is once removed from ω and write as $\omega\mathsf{R}^n\omega'$.

Proposition 1 *If ω is a world in a standard model $\mathfrak{M} = (W, \mathsf{R}, \nu)$, then for $n \geq 0$:*

- $\mathfrak{M}, \omega \models \Box^n \varphi$ *iff for every* ω' *in* $\mathfrak{M}$ *such that* $\omega \mathsf{R}^n \omega'$, $\mathfrak{M}, \omega' \models \varphi$
- $\mathfrak{M}, \omega \models \Diamond^n \varphi$ *iff for some* ω' *in* $\mathfrak{M}$ *such that* $\omega \mathsf{R}^n \omega'$, $\mathfrak{M}, \omega' \models \varphi$

The result of this proposition could be generalised with the $G^{k,l,m,n}$ schema as mentioned above and we could see that the schemas **D, T, B, 4, 5** as shown in Figure 2.1 are special cases of $G^{k,l,m,n}$.

$$G^{k,l,m,n} = \Diamond^k \Box^l \varphi \Rightarrow \Box^m \Diamond^n \varphi \quad \begin{cases} \mathbf{D} = G^{0,1,0,1} \\ \mathbf{T} = G^{0,1,0,0} \\ \mathbf{B} = G^{0,0,1,1} \\ \mathbf{4} = G^{0,1,2,0} \\ \mathbf{5} = G^{1,0,1,1} \end{cases}$$

Theorem 6 *If $\mathfrak{M}^\Lambda_\zeta = (\mathsf{W}^\Lambda_\zeta, \mathsf{R}^\Lambda_\zeta, \nu^\Lambda_\zeta)$ be the canonical model for a normal logic Λ, then for every ω,ω' in $\mathfrak{M}$ and every $k \geq 0$*

- $\omega \mathsf{R}^k \omega'$ *iff* $\{\varphi : \Box^k \varphi \in \omega \} \subseteq \omega'$
- $\omega \mathsf{R}^k \omega'$ *iff* $\{\Diamond^k \varphi : \varphi \in \omega' \subseteq \omega$

Theorem 7 *The schema $G^{k,l,m,n}$ is valid in the class of k,l,m,n-incestual models.*

From this theorem we can conclude that the modal logic $\mathbf{K} \oplus G^{k,l,m,n}$ is sound with respect to any class of k, l, m, n-incestual models. The completeness of the logic is stated in the following theorem the proof of which could be found in [22].

Theorem 8 *The canonical model for a normal $\mathbf{K} \oplus G^{k,l,m,n}$ logic is k,l,m,n-incestual, for every $k, l, m, n \geq 0$.*

2.2 Multi-Modal Logics

Multimodal logics generalise modal logics allowing more than one modal operator to appear in formulae. They are particularly suitable to reason in a multiagent environment, to modal several agents and to represent group properties like *knowledge*, *beliefs* and flow of time. For instance, if one wants to represent the beliefs of n agents developing in time one may need

n+1 pairs of boxes and diamonds where one pair is used to talk about time and one pair for representing the beliefs of each agent. The only difference between $\mathcal{L}_{\mathbf{PML}}$ (propositional modal language) as given in the previous section and $\mathcal{L}_{\mathbf{PML}_n}$ (propositional multi-modal language) is that $\Box_i\varphi$ and $\Diamond_i\varphi$ are formulas of $\mathcal{L}_{\mathbf{PML}_n}$ whenever $1 \leq i \leq n$ and φ is an $\mathcal{L}_{\mathbf{PML}_n}$ formula (the difference amounts to have $\Box_1, \ldots, \Box_n$ necessity operators and $\Diamond_1, \ldots, \Diamond_n$ possibility operators). The syntax of n-modal normal logics consists of the **K**-axiom and the generalisation rule formulated for each of the boxes $\Box_1, \ldots, \Box_n$

- (**K**) $\Box_i(\mathfrak{p}_0 \Rightarrow \mathfrak{p}_1) \Rightarrow (\Box_i\mathfrak{p}_0 \Rightarrow \Box_i\mathfrak{p}_1)$
- (NEC) given φ, derive $\Box_i\varphi$

A set of $\Lambda_{\mathbf{PML}_n}$-formulas is called an *n-modal logic* if it contains all propositional tautologies and (**K**), for $1 \leq i \leq n$, and is closed under the rules modus ponens (MP), substitution (Subst) and Necessitation (NEC), for all $i = 1, \ldots, n$. $\mathbf{K}_n$ is defined as the smallest (minimal) *n-modal logic*. As before, for afn n-modal Λ_0 and a set Σ of $\mathcal{L}_{\mathbf{PML}_n}$ formulas, the smallest n-modal logic is denoted by $\Lambda_0 \oplus \Sigma$ containing $\Lambda_0 \cup \Sigma$. Table 2.4 shows examples of multi-modal logics for n-modal variants of some logics.

Name of the Logic	Axiom
$\mathbf{K4}_n$	$\mathbf{K}_n \oplus \{\Box_i\mathfrak{p}_0 \Rightarrow \Box_i\Box_i\mathfrak{p}_0 \mid 1 \leq i \leq n\}$
$\mathbf{T}_n$	$\mathbf{K}_n \oplus \{\Box_i\mathfrak{p}_0 \Rightarrow \mathfrak{p}_0 \mid 1 \leq i \leq n\}$
$\mathbf{S4}_n$	$\mathbf{K4}_n \oplus \{\Box_i\mathfrak{p}_0 \Rightarrow \mathfrak{p}_0 \mid 1 \leq i \leq n\}$
$\mathbf{KD45}_n$	$\mathbf{K4}_n \oplus \{\Box_i\mathfrak{p}_0 \Rightarrow \Diamond_i\mathfrak{p}_0, \Diamond_i\mathfrak{p}_0 \Rightarrow \Box_i\Diamond_i\mathfrak{p}_0 \mid 1 \leq i \leq n\}$
$\mathbf{S5}_n$	$\mathbf{S4}_n \oplus \{\Diamond_i\mathfrak{p}_0 \Rightarrow \Box_i\Diamond_i\mathfrak{p}_0 \mid 1 \leq i \leq n\}$

Table 2.4: Examples of Multi-Modal Logics

It should be noted that the axioms of the different multi-modal logics shown in Table 2.4 does not allow any interaction between the different modal operators. Therefore, the axioms of $\mathbf{K4}_n$ require each $\Box_i$ to behave like a **K4**-box and the same is the case with the other logics too. No axiom with two different boxes is shown and hence there is no interaction. As we will show in the next chapter this *interaction* between the modal operators forms a major part of this thesis.

The interpretation of $\Lambda_{\mathbf{PML}_n}$-formulas in terms of possible worlds semantics requires n accessibility relations $\mathsf{R}_1, \ldots, \mathsf{R}_n$ (one for each $\Box_i$) between worlds in a Kripke frame (W, R). Thus the structure of an *n-frame* has the form

$$\mathcal{F} = (\mathsf{W}, \mathsf{R}_1, \ldots, \mathsf{R}_n)$$

where W is a non-empty set of of worlds and $\mathsf{R}_1, \ldots, \mathsf{R}_n$ is the n binary relations on W. *Valuation* in an *n-frame* $\mathcal{F}$ is as usual a map ν associating with each propositional variable $\mathfrak{p}_0$ a subset $\nu(\mathfrak{p}_0)$ of W. The pair $\mathfrak{M} = (\mathcal{F}, \nu)$ is a *model* for Λ_{PML_n}. Satisfaction is defined as follows:

$$\mathfrak{M}, \omega \models \Box_i\varphi \text{ iff } (\mathfrak{M}, \omega') \models \varphi \text{ for all } \omega' \in \mathsf{W} \text{ such that } \omega \mathsf{R}_i \omega'$$
$$\mathfrak{M}, \omega \models \Diamond_i\varphi \text{ iff } (\mathfrak{M}, \omega') \models \varphi \text{ for some } \omega' \in \mathsf{W} \text{ such that } \omega \mathsf{R}_i \omega'$$

for all $i = 1, \ldots, n$. Given a class $\mathfrak{F}$ of *n-frames* the logic of $\mathfrak{F}$ is defined as

$$\Delta_{\mathfrak{F}} = \{\varphi \in \mathcal{L}_{\mathbf{PML}_n} \mid \forall \mathcal{F} \in \Delta\mathcal{F} \models \varphi\} \tag{2.3}$$

The only difference between this formulation and the one given in (2.1) is with regard to the n modalities. In a similar manner the definitions given in the previous section for the monomodal case could be interpreted for the multi-modal language $\mathcal{L}_{\mathbf{PML}_n}$.

Theorem 9 *The n-modal logics* $\mathbf{K}_n$, $\mathbf{K4}_n$, $\mathbf{T}_n$, $\mathbf{S4}_n$, $\mathbf{KD45}_n$ *and* $\mathbf{S5}_n$ *are complete with respect to their n-frames as given in Table 2.4*

2.2.1 BDI Logics

In this section we give an overview of how a modal language as developed in the previous sections can be used to represent the mental states of an agent, in particular its *belief-desire-intention* state (BDI-model). Here we describe only the static fragment of BDI logics, so no temporal evolution will be present (for further reference see [101]). The language ($\mathcal{L}_{\mathbf{BDI}}$) is a propositional modal language with three families of modal operators $\mathrm{BEL}_i, \mathrm{DES}_i, \mathrm{INT}_i$, $i \in A$ (agents) representing, respectively, the beliefs, desires and intentions of the agents. The **K**-axiom of modal logic is adopted for beliefs, desires and intentions. As mentioned earlier the **K**-axiom characterize the minimal system for NML's. This axiom states that if an agent believes φ and and believes $\varphi \Rightarrow \psi$ then he/she will believe ψ. This constraint is extended to desires and intentions. In addition to the **K**-axiom, axioms **D, 4, 5** are adopted for beliefs which gives rise to the logic **KD45** which is basically the logic **S5** in which the axiom **T**:$\Box(\varphi) \Rightarrow \varphi$ is replaced by the weaker **D**:$\Box\varphi \Rightarrow \neg\Box\neg\varphi$. The reason for this change of axiom results from the fact that the *belief* of an agent may be false as opposed to its *knowledge* which cannot be false. Hence, the **D**-axiom expresses the consistency of beliefs and the **4**-axiom and **5**-axiom expresses the positive and

negative introspection capabilities of an agent with respect to its beliefs. The logic **KD45** can be axiomatised as shown below where BEL(φ) means *the agent believes that* φ.

(**K**) BEL($\varphi \Rightarrow \psi$) $\Rightarrow$ (BEL(φ) $\Rightarrow$ BEL(ψ)
(**D**) BEL(φ) $\Rightarrow \neg$BEL$\neg(\varphi)$
(**4**) BEL(φ) $\Rightarrow$ BEL(BEL(φ))
(**5**) $\neg$BEL(φ) $\Rightarrow$ BEL($\neg$BEL(φ))
($Taut$) $\top$ where $\top$ is any propositional tautology
(US) If φ, then $\varphi\{\psi_1/p_1 \ldots \psi_n/p_n\}$
(MP) If φ and $\varphi \Rightarrow \psi$, then ψ
(Nec) If φ, then BEL(ψ)

For desires and intentions, in addition to the **K**-axiom we adopt the standard **D**-axiom which expresses the consistency of desires and intentions. We have the following axioms for desires and intentions

(**K**) INT($\varphi \Rightarrow \psi$) $\Rightarrow$ (INT(φ) $\Rightarrow$ INT(ψ)
(**K**) DES($\varphi \Rightarrow \psi$) $\Rightarrow$ (DES(φ) $\Rightarrow$ DES(ψ)
(D) DES(φ) $\Rightarrow \neg$DES$\neg(\varphi)$
(D) INT(φ) $\Rightarrow \neg$INT$\neg(\varphi)$
($Taut$) $\top$ where $\top$ is any propositional tautology
(US) If φ, then $\varphi\{\psi_1 \ldots \psi_n/p_n\}$
(MP) If φ and $\varphi \Rightarrow \psi, then \psi$
(Nec) If φ, then INT(φ)
(Nec) If φ, then DES(φ)

The semantics is given through Kripke structures which is explained as follows

Definition 13 *A Kripke structure is defined to be a tuple*

$$M = (\mathsf{W}, \{S_\omega : \omega \in \mathsf{W}\}, \{\mathsf{R}_\omega : \omega \in \mathsf{W}\}, \nu, \mathsf{B}, \mathsf{G}, \mathsf{I}),$$

where W *is a set of possible worlds,* S_ω *is the set of states in each world* W*,* R_ω *is a binary relation, i.e.,* $\mathsf{R}_\omega \subseteq S_\omega \times S_\omega, \nu$ *is a truth assignment to the primitive propositions of* Φ *for each world* $\omega \in \mathsf{W}$ *at each state* $s \in S_\omega$*, (i.e.,* $\nu(\omega, s) : \Phi \rightarrow \{$ *true, false* $\}$*, and* B*,* G *and* I *are relations on the worlds,* W *and states,* S *(i.e.,* $\mathsf{B} \subseteq \mathsf{W} \times S \times \mathsf{W}$*).*

Definition 14 *A world* ω' *is a* ***sub-world*** *of the world* ω*, denoted* $\omega' \sqsubseteq \omega$*, if and only if*

- $S_{\omega'} \subseteq S_\omega$

- $\mathsf{R}_{\omega'} \subseteq \mathsf{R}_{\omega}$
- $\forall s \in S_{\omega'}$, $\nu(\omega', s) = \nu(\omega, s)$
- $\forall s \in S_{\omega'}, (\omega', s, \omega'') \in \mathsf{B}$ *iff* $(\omega', s, \omega'') \in \mathsf{B}$*; and similarly for* G *and* I.

Definition 15 *ω' is a **strict sub-world** of ω, denoted by $\omega' \sqsubset \omega$, if and only if $\omega' \sqsubseteq \omega$ and $\omega \not\sqsubseteq \omega'$. If ω' is a sub-world of ω then ω is a **super-world** of ω', denoted by $\omega \sqsupseteq \omega'$. ω' is said to be **structurally equivalent** to ω, denoted by $\omega' \leftrightharpoons \omega$ iff $\omega' \sqsubseteq \omega$ and $\omega \sqsubseteq \omega'$.*

It should be noted that the above definitions correspond to a BDI-model in a temporal setting as given in the original work of Rao and Georgeff [101]. But as pointed out earlier, in this thesis we do not make any reference to temporal notions. We use the above definition to differentiate the *set* and *structural* relationship that exists between BEL, GOAL and INT modalities as will be shown in the next section. Since in this work we focus on the set relationship we need to augment the above definition appropriately. We define a Kripke structure to be a tuple

$$\mathfrak{M} = (\mathsf{W}, \nu, \mathsf{B}, \mathsf{G}, \mathsf{I}),$$

where W is a set of possible worlds, ν is a truth assignment to the primitive propositions of Φ for each world $\omega \in \mathsf{W}$ (i.e., $\nu(\omega) : \Phi \rightarrow \{true, false\}$, and BEL, GOAL and INT are binary relations on the worlds of W (i.e., BEL $\subseteq \mathsf{W} \times \mathsf{W}$). Satisfaction of formulas is given with respect to a structure $\mathfrak{M}$, and a world ω. The expression $\mathfrak{M}, \omega \models \varphi$ is read as *structure $\mathfrak{M}$ at world ω satisfies φ.*

$\mathfrak{M}, \omega \models \mathfrak{p}_0$ iff $\omega \in \nu(\mathfrak{p}_0)$ where $\mathfrak{p}_0$ is a primitive proposition ;
$\mathfrak{M}, \omega \models \neg\varphi$ iff $\mathfrak{M}, \omega \not\models \varphi$;
$\mathfrak{M}, \omega \models \varphi \wedge \psi$ iff $\mathfrak{M}, \omega \models \varphi$ and $\mathfrak{M}, \omega \models \psi$;
$\mathfrak{M}, \omega \models \text{BEL}(\varphi)$ iff $\forall\omega'$ satisfying $(\omega, \omega') \in \mathsf{B}, \mathfrak{M}, \omega' \models \varphi$;
$\mathfrak{M}, \omega \models \text{GOAL}(\varphi)$ iff $\forall\omega'$ satisfying $(\omega, \omega') \in \mathsf{G}, \mathfrak{M}, \omega' \models \varphi$;
$\mathfrak{M}, \omega \models \text{INT}(\varphi)$ iff $\forall\omega'$ satisfying $(\omega, \omega') \in \mathsf{I}, \mathfrak{M}, \omega' \models \varphi$.

We say that an agent has a belief φ (BEL(φ)) in world s iff φ is true in all the belief-accessible worlds of the agent. As the belief-accessibility relation is dependent on the world, the mapping of BEL at some other world may be different. The reason for having multiple belief-accessible worlds is to allow different worlds (possibilities) to be modeled in cases where the agent lacks enough information (or is uncertain) about the current

world. Similar to belief-accessible worlds, there is also a set of *desire (goal)-accessible* worlds to represent the desires (goals) of the agent. An agent has a desire φ in world ω iff φ is true in all the desire (goal)-accessible worlds of the agent. Intentions are also modeled in this way using an intention-accessibility relation. Intention worlds are the ones the agent has *chosen* to attempt to realise. We say that the agent intends a formula in a certain world iff it is true in all the agent's intention-accessible worlds.

Validity is defined in the standard manner where a formula is valid if it is true in every world of every structure. A formula φ is said to be *valid* in $\mathfrak{M}$, written as $\mathfrak{M} \models \varphi$, if $\mathfrak{M}, \omega \models \varphi$ for every world $\omega \in \mathsf{W}$. Similarly, validity and satisfiability with respect to a class $\mathcal{M}$ of structures can be defined. φ is valid with respect to a class $\mathcal{M}$ of structures ($\mathcal{M} \models \varphi$) if φ is valid in all structures in $\mathcal{M}$ and φ is satisfiable with respect to a class $\mathcal{M}$ of structures if φ is satisfiable in some structure in $\mathcal{M}$.

Theorem 10 *[100] For all formulas $\varphi, \psi \in \mathcal{L}_{BDI}$, structures $\mathfrak{M} \in \mathcal{M}$,*

1. *if φ is an instance of a propositional tautology, then $\mathfrak{M} \models \varphi$;*
2. *if $\mathfrak{M} \models \varphi$ and $\mathfrak{M} \models \varphi \Rightarrow \psi$, then $\mathfrak{M} \models \psi$;*
3. *$\mathfrak{M} \models (\Box(\varphi) \wedge \Box(\varphi \Rightarrow \psi)) \Rightarrow \Box(\psi)$, where $\Box$ stands for* BEL*,* GOAL*,* INT*;*
4. *if $\mathfrak{M} \models \varphi$ then $\mathfrak{M} \models \Box(\varphi)$, where $\Box$ stands for* BEL*,* GOAL *and* INT*.*

Proof. 1 and 2 follows from from **PL**. For part 3 the proof runs as follows: If $\mathfrak{M}, \omega \models \Box(\varphi) \wedge \Box(\varphi \Rightarrow \psi)$, then for all worlds ω' such that $(\omega, \omega') \in \Box$, we have $\mathfrak{M}, \omega' \models \varphi$ and $\mathfrak{M}, \omega' \models \varphi \Rightarrow \psi$. By propositional reasoning we have $\mathfrak{M}, \omega' \models \psi$ for all ω' such that $(\omega, \omega') \in \Box$. Hence, $\mathfrak{M}, \omega \models \Box(\psi)$. As this line of reasoning holds for any arbitrary $\omega \in \mathsf{W}$, we have $\mathfrak{M} \models (\Box(\varphi) \wedge \Box(\varphi \Rightarrow \psi)) \Rightarrow \Box(\psi)$.

If $\mathfrak{M} \models \varphi$ then $\mathfrak{M}, \omega' \models \varphi$ for all worlds ω' in $\mathfrak{M}$. In particular if we consider a world ω, then $\mathfrak{M}, \omega' \models \varphi$ for all ω' such that $(\omega, \omega') \in \Box$. Thus, $\mathfrak{M}, \omega \models \varphi$ for all worlds ω in $\mathfrak{M}$, and hence $\mathfrak{M} \models \Box(\varphi)$ □

Since we will introduce a number of different logics it is recommended to use a common nomenclature for identifying them. As shown above beliefs, desires and intentions are abbreviated as BEL, DES and INT and their respective axiom systems are written as a superscript. Thus the modal logic $\text{BEL}^{(\mathbf{K})}, \text{DES}^{(\mathbf{K})}$ and $\text{INT}^{(\mathbf{K})}$ shows that we have adopted the **K**-axiom system for beliefs, desires and intentions. When the same axiom system is used for the modal operators BEL, DES and INT the notation is simplified

by writing the superscript once for all the modal operators as in $\text{BDI}^{(\mathbf{K})}$. According to this nomenclature the above axiom systems for beliefs, desires and intentions is denoted by $\mathsf{B}^{(\mathbf{KD45})}, \mathsf{D}^{(\mathbf{KD})}, \mathsf{I}^{(\mathbf{KD})}$ (as noted in the introduction, for the purpose of this thesis we don't differentiate between DES and GOAL and hence the rules remain the same for both of them).

Lemma 7 *[64] In any axiom system Λ that includes all propositional tautologies and modus ponens, every consistent set Δ can be extended to a maximal Λ-consistent set. In addition, if Δ is a maximal consistent set, then it satisfies the following properties:*

1. *for every formula φ, exactly one of φ and $\neg\varphi$ is in Δ;*
2. *$\varphi \wedge \psi \in \Delta$ iff $\varphi \in \Delta$ and $\psi \in \Delta$;*
3. *if φ and ψ are both in Δ, then ψ is in Δ;*
4. *if φ is $(\text{BDI})^{\mathbf{K}}$-provable, then $\varphi \in \Delta$.*

Using the standard definitions of soundness and completeness as given in the previous section and the Lemma given above Rao proves the following two theorems:

Theorem 11 *[100] The basic axiom system $(\text{BDI})^{\mathbf{K}}$ is a sound and complete axiomatisation with respect to the unrestricted class of structures $\mathcal{M}$.*

Theorem 12 *[100](**Soundness and completeness of basic BDI systems**)*

1. $\text{BDI}^{(\boldsymbol{T})}$ *is a sound and complete axiomatisation with respect to a class $\mathcal{M}$ of structures where the relations B, D and I are reflexive;*
2. $\text{BDI}^{(\boldsymbol{S4})}$ *is a sound and complete axiomatisation with respect to a class $\mathcal{M}$ of structures where the relations B, D and I are reflexive and transitive;*
3. $\text{BDI}^{(\boldsymbol{S5})}$ *is a sound and complete axiomatisation with respect to a class $\mathcal{M}$ of structures where the relations B, D and I are reflexive, symmetric and transitive;*
4. $\text{BDI}^{(\mathbf{KD45})}$ *is a sound and complete axiomatisation with respect to a class $\mathcal{M}$ of structures where the relations B, D and I are Euclidean, serial and transitive;*

5. $B^{(\mathbf{KD45})}D^{(\boldsymbol{KD})}I^{(\boldsymbol{T})}$ *is a sound and complete axiomatisation with respect to a class* $\mathcal{M}$ *of structures where the relation* B *is Euclidean, serial and transitive;* D *is serial and* I *is reflexive;*

6. $B^{(\mathbf{KD45})}D^{(\boldsymbol{S5})}I^{(\boldsymbol{S5})}$ *is a sound and complete axiomatisation with respect to a class* $\mathcal{M}$ *of structures where the relation* B *is Euclidean, serial and transitive;* D *and* I *are reflexive, symmetric and transitive.*

2.2.2 BDI logics with interaction axioms

The axiom systems of beliefs, desires and intentions considered so far do not capture the rich inter-relationships between the different mental modalities. In this section we outline the various multi-modal axioms that capture these relationships and also show the corresponding semantic conditions. But, before giving the semantic conditions and the corresponding axioms for the various relationships among the mental modalities (for BDI) we give an overview of other related work that has been done to capture the relationship between beliefs, goals and intentions.

Most of the formalisms dealing with the inter-relationship of mental modalities has roots in the philosophical work of Bratman [12]. Bratman argues that it is irrational for an agent to intend to do an action and also believe that he will not do it which is formally given as

$$\not\models \mathrm{INT}(\varphi) \wedge \mathrm{BEL}(\neg\varphi) \tag{2.4}$$

At the same time it is rational for the agent to intend to do an action φ but not believe that he will do it.

$$\models \mathrm{INT}(\varphi) \wedge \neg\mathrm{BEL}(\varphi) \tag{2.5}$$

The difference here is that, it is irrational for an agent to have beliefs that are inconsistent with his intentions but perfectly rational to have incomplete beliefs about his intentions. These two principles of *intention-belief consistency* (2.4) and *intention-belief incompleteness* (2.5) has been termed as *asymmetry thesis* by Bratman. The *asymmetry thesis* could be extended to intentions and goals and goals and beliefs maintaining *intention-goal consistency* and *goal-belief consistency* as well as *intention-goal incompleteness* and *goal-belief incompleteness.* The way in which the relationships between beliefs, goals and intentions are captured can have a significant impact on the design of a rational agent. Two related problems that can arise while defining the inter-relationships between the different mental modalities is that of *side-effect problem* and *transference problem.* The side-effect problem (in relation to belief and intention) states that an agent who intends

to do φ should not be forced to do ψ, no matter how strongly[5] he believes that doing φ will force him to do ψ. Formally this is given as

$$\models \text{INT}(\varphi) \wedge \Box\text{BEL}(\Box(\varphi \Rightarrow \psi)) \wedge \neg\text{INT}(\psi) \tag{2.6}$$

On the other hand the problem of transference states that no matter how strongly an agent believes in a proposition, he should not be forced to adopt it as a goal and is formally given as

$$\models \text{BEL}(\varphi) \wedge \neg\text{GOAL}(\varphi) \tag{2.7}$$

The four conditions given above are called *principles* by Bratman and for him a rational agent is one who satisfies all the above principles.

The first direct attempt to provide a logical analysis of Bratman's theory as explained above was made by Cohen and Levesque [25, 24], though Allen was the first one to give a formal notion of intention [2]. But for Allen intentions could be reducible to beliefs about future actions. According to Cohen and Levesque's formalism the semantics of BEL and DES are given by belief- and goal(des)-accessibility relations on possible worlds where goal(des)-accessibility worlds are a subset of belief-accessibility worlds, (i.e., $\mathsf{G} \subseteq \mathsf{B}$). Each possible world in their formalism is a time-line that imposes the condition that the chosen (or desire-accessible worlds) be compatible with the agent's belief-accessible worlds. This is called the *realism constraint* and it is meant to ensure that the worlds chosen by the agent are not ruled out by its beliefs. Figure 2.3 shows the subset (Fig (A)) and sub-world (Fig (B)) relations. Formally this could be shown as:

$\text{BEL}(\mathfrak{p}_0) \Rightarrow \text{GOAL}(\mathfrak{p}_0)$
$\forall\omega\forall\omega'$ if $(\omega, \omega') \in \mathsf{G}$ then $(\omega, \omega') \in \mathsf{B}$ (or GOAL $\subseteq$ BEL)

$\text{GOAL}(\mathfrak{p}_0) \Rightarrow \text{INT}(\mathfrak{p}_0)$
$\forall\omega\forall\omega'$ if $(\omega, \omega') \in \mathsf{I}$ then $(\omega, \omega') \in \mathsf{G}$ (or INT $\subseteq$ GOAL)

which means *if an agent i believes a proposition $\mathfrak{p}_0$ it will also have the intention (or goal) towards that proposition.* These axioms correspond to a *multi-modal containment condition* where all intention-accessible worlds are contained in the set of goal-accessible worlds and all goal-accessible worlds are contained in the set of belief-accessible worlds. This relationship is shown in Fig. 2.3.

Based on this proposition Cohen and Levesque go on to define what they call *achievement goal, persistent goal, relativised persistent goal etc..*

[5]The three types of strong conditions given by Cohen and Levesque are $\text{BEL}(\varphi \Rightarrow \psi)$, $\text{BEL}(\Box(\varphi \Rightarrow \psi))$ or $\Box\text{BEL}(\Box(\varphi \Rightarrow \psi))$.

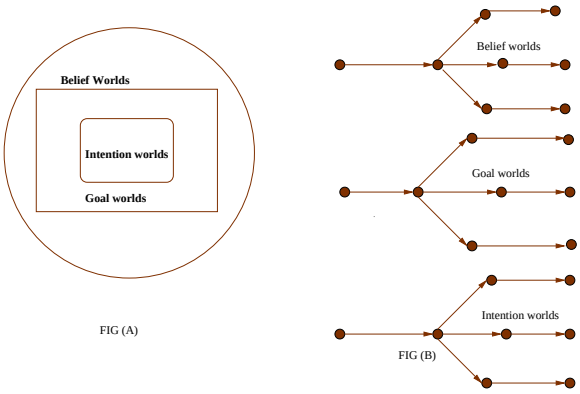


Figure 2.3: Realism possible worlds structure ($\mathsf{I} \subseteq \mathsf{G} \subseteq \mathsf{B}$)

For a good overview refer [63]. Several criticisms have been raised against Cohen and Levesque's work [109, 105, 116] of which the work by Rao and Georgeff [105] is important as far as this thesis is concerned. Rao and Georgeff argue that though the above axioms ensure that the goals of an agent are not ruled out by its beliefs, it forces the agent to adopt as beliefs certain inevitable facts about the world. For them, the realism constraint characterises an agent that intends all future propositions that it desires (or has as its goals) to bring about and desires all future propositions that it *believes* can be achieved. Such an agent is *over-enthusiastic*. Hence they introduce the notion of *strong realism*, which is captured by imposing the following restrictions.

$\text{GOAL}(\varphi) \Rightarrow \text{BEL}(\varphi)$
$\forall\omega, \forall\omega'$ if $(\omega, \omega') \in \mathsf{B}$ then $\exists\omega'', (\omega, \omega'') \in \mathsf{G}$ and $\omega' \sqsupseteq \omega''(or\ \mathsf{B} \subseteq_{sup} \mathsf{G})$

The constraint states that for each and every belief-accessible world there is a corresponding goal-accessible world such that the goal-world is a *sub-world* of the belief-world. In other words, a *strong realism constraint* is one where the set of belief-accessible worlds is a *subset* of goal-accessible worlds and each belief- accessible world is a super-world of some goal-accessible world ($\mathsf{B}^{\omega} \subseteq_{sup} \mathsf{G}^{\omega}$). The notion of subset and sub-world\super-world relationship between belief- and goal-accessible worlds arises due to a difference in the set and structural relationship among the possible worlds. In [107] the relationship between the belief-, desire-, and intention- accessible worlds are examined with respect to the *set relationship* among the possible worlds as well as with respect to the *structure* of possible worlds. For instance, given two sets Σ and Γ, the following relationships can hold between them: $\Sigma \subseteq \Gamma$, $\Gamma \subseteq \Sigma$, $\Sigma \cap \Gamma \neq \emptyset$, and $\Sigma \cap \Gamma = \emptyset$. These set relationships hold between

the sets of belief- and goal(desire)-accessible worlds, goal- and intention-accessible worlds and belief- and intention-accessible worlds. As far as the structural relationships are concerned since each possible world is a time tree, one can consider additional structural relationships between two given worlds. Given two worlds ω and ω', if the tree structure of ω' is a sub-tree of ω and has the same truth-assignment and accessibility relations as ω then ω' is a sub-world of ω. Hence if ω and ω' are two worlds then

- ω could be a sub-world of ω';
- ω' could be a sub-world of ω;
- ω and ω' could be identical;
- ω and ω' could be totally different.

The set and structural relationships can be combined to obtain a variety of different possible-world structures. Figure 2.4(A) shows the subset relation ship for strong-realism whereas Figure 2.4(B) indicates the sub-world relation for the same.

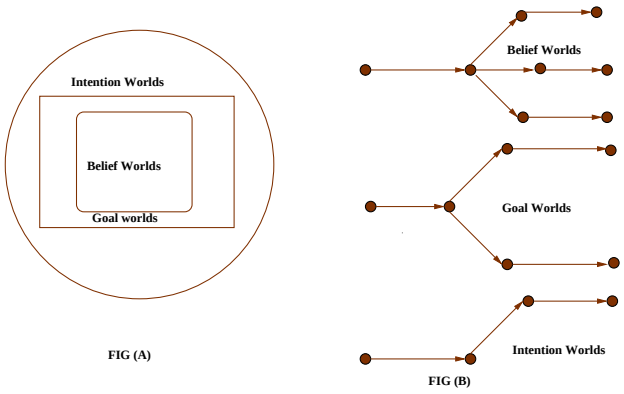


Figure 2.4: Strong realism possible worlds structure ($\mathsf{B}^{\omega} \subseteq_{sup} \mathsf{G}^{\omega} \subseteq_{sup} \mathsf{I}^{\omega}$)

The strong-realism axiom mentioned above holds when φ is an **O**-formula. Well-formed formulas that contain no positive occurrences of *inevitable* (or negative occurrences of *optional*)[6] outside the scope of the modal operators BEL, GOAL or INT are called **O**-formulas in [107]. The

[6] *inevitable* and *optional* are two modalities which operate on *path formulas*. Rao and Georgeff uses Computation Tree Logic (CTL^* [32, 33]) to distinguish between *state* and *path* formulas where the former are evaluated at a specified time point in a time tree and the latter over a specified path in a time tree.

axiom states that if the agent has the goal (desire) towards *optional(φ)*, it also believes that *optional(φ)*; or in other words there is at-least one path in all the belief-accessible worlds in which φ is true. In a similar manner, if the formula ψ above is *eventually* $\mathfrak{p}_0$ (sometime in the future $\mathfrak{p}_0$), then the axiom states that in all the belief-accessible worlds of the agent there is at-least one path where *eventually* $\mathfrak{p}_0$ is true. But, because of the branching nature of time, the agent need not believe that it will ever reach the state where $\mathfrak{p}_0$ is true. The semantic condition for strong realism in terms of GOAL and INT can be given in similar lines as:

$\text{INT}(\varphi) \Rightarrow \text{GOAL}(\varphi)$
$\forall\omega, \forall\omega'$ if $(\omega, \omega') \in \mathsf{G}$ then $\exists\omega'', (\omega, \omega'') \in \mathsf{I}$ and $\omega' \sqsupseteq \omega''(or\ \mathsf{G} \subseteq_{sup} \mathsf{I})$

There is a slight difference in the interpretation of **I**-formulas[7] in terms of the sub-set and structural relationships. In the case of **I**-formulas the subset relationship is kept the same as in the case of **O**-formulas, whereas in the structural relationship a belief-accessible world is a sub-world of the goal-accessible world which in turn is a sub-world of an intention-accessible world. Hence the semantic condition runs like:

$\text{GOAL}(\chi) \Rightarrow \text{BEL}(\chi)$
$\forall\omega\forall\omega'$ if $(\omega, \omega') \in \mathsf{B}$ then $\exists\omega'', (\omega, \omega'') \in \mathsf{G}$ and $\omega' \sqsubseteq \omega''(\mathsf{B} \subseteq_{sub} \mathsf{G})$

$\text{INT}(\chi) \Rightarrow \text{GOAL}(\chi)$
$\forall\omega\forall\omega'$ if $(\omega, \omega') \in \mathsf{G}$ then $\exists\omega'', (\omega, \omega'') \in \mathsf{I}$ and $\omega' \sqsubseteq \omega''(\mathsf{G} \subseteq_{sub} \mathsf{I})$

If the structures in corresponding belief-, goal- and intention-accessible worlds are identical then the semantic condition can be given as:

$\text{GOAL}(\psi) \Rightarrow \text{BEL}(\psi)$
$\forall\omega\forall\omega'$ if $(\omega, \omega') \in \mathsf{B}$ then $(\omega, \omega') \in \mathsf{G}$ $(\mathsf{B} \subseteq \mathsf{G})$

$\text{INT}(\psi) \Rightarrow \text{GOAL}(\psi)$
$\forall\omega\forall\omega'$ if $(\omega, \omega') \in \mathsf{G}$ then $(\omega, \omega') \in \mathsf{I}$ $(\text{GOAL} \subseteq \text{INT})$

It could also be said that these axioms correspond to a *multi-modal containment condition* where all goal-accessible worlds are contained in the set of intention-accessible worlds and all belief-accessible worlds are contained in the set of goal-accessible worlds. As a result of this if the agent desires to *optionally* achieve a proposition, the agent also believes the proposition to

[7] **I**-formulas do not contain any positive occurrences of *optional* outside the scope of the modal, operators BEL, GOAL and INT.

be an option it can achieve (if it chooses). Under strong realism, different belief-, desire- and intention accessible worlds represent different possible scenarios for the agent. Intuitively, the agent believes the actual world to be one of its belief-accessible worlds; if it were to be in belief world b_1, then its goals (with respect to b_1) would be a corresponding goal-accessible world, g_1, and its intentions a corresponding intention-accessible world i_1. The worlds g_1 and i_1 represent increasingly selective choices from b_1 about the goal and choice of possible future courses of action. But this doesn't mean that rational agents based on strong-realism constraints gives the best choice. It again depends on the domain of choice.

Rational agents based on the strong-realism constraint are *over-cautious* in that they only desire future propositions that are believed to be achievable and only intend future propositions that are part of their desires. Though such constraints form an important part in the design of agents with mental states they do not seem to be in agreement with the sort of reasoning in human beings. A balance between the two can be obtained if agents have the property that they do not desire propositions the negations of which are believed; do not intend propositions the negations of which are desired; and do not intend propositions the negations of which are believed by the agent. Such a property is called *weak realism* [105] and is formally given as:

$$\text{INT}(\mathfrak{p}_0) \Rightarrow \neg\text{GOAL}(\neg\mathfrak{p}_0)$$
$$\forall\omega\forall\omega'\exists\omega'' \ (\omega,\omega',\omega'') \ \in \mathsf{I} \text{ iff } (\omega,\omega',\omega'') \in \mathsf{G} \text{ (or } \mathsf{G}\cap\mathsf{I}\neq\emptyset)$$

$$\text{INT}(\mathfrak{p}_1) \Rightarrow \neg\text{BEL}(\neg\mathfrak{p}_1)$$
$$\forall\omega\forall\omega'\exists\omega'' \ (\omega,\omega',\omega'') \ \in \mathsf{I} \text{ iff } (\omega,\omega',\omega'') \in \mathsf{B} \text{ (or } \mathsf{B}\cap\mathsf{I}\neq\emptyset)$$

$$\text{GOAL}(\mathfrak{p}_2) \Rightarrow \neg\text{BEL}(\neg\mathfrak{p}_2)$$
$$\forall\omega\forall\omega'\exists\omega'' \ (\omega,\omega',\omega'') \ \in \mathsf{G} \text{ iff } (\omega,\omega',\omega'') \in \mathsf{B} \text{ (or } \mathsf{B}\cap\mathsf{G}\neq\emptyset)$$

These axioms correspond to a *multi-modal version of the seriality condition* which semantically requires that there is at least one world common to intention-accessible worlds and belief-accessible worlds. Similarly, it is also required that there should be at least one world that is both intention-accessible and goal-accessible and also at least one world that is both goal-accessible and belief-accessible. Figure 2.5 depicts this relationship.

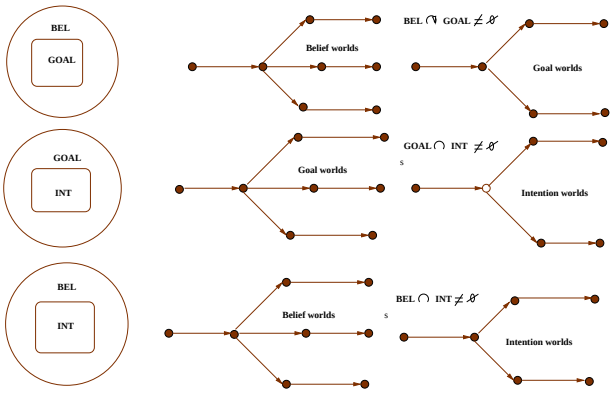


Figure 2.5: Weak realism possible worlds structure

2.3 Combining Logics

As shown in the previous section, BDI-logics are multi-modal logics for modeling belief, desire (goal) and intention; comprises of three different logical layers (**KD45** for belief, **KD** for goals (desires) and **KD** for intentions) and is linked together by specific interaction axioms. BDI logics are nothing but extensions of many basic normal logics *combined together* to model different facets of the agents. This combined nature of BDI-logics makes it worthwhile in exploring certain existing combining techniques in the background of BDI and thereby prove some general results about them. We explore two such techniques, *fusion* and *fibring*.

The field of *combining logics* is emerging as an active area, promising powerful results about the preservation of important properties of the logics being combined [77, 42, 43, 114, 128, 129]. The novelty of combining logics is the aim to develop *general techniques* that allow us to produce combinations of *existing* and well understood logics. Such general techniques provide answer to the need of formalising complex systems in a systematic way. Combining logics recognizes that the big problems involved in formalising complex logics such as those required in many areas of AI and Linguistics require two main strategies: *divide and conquer* and *hide and unpack*[8], i.e., try to split problems and delegate sub-problems to the component logics and when working inside one of the component logics view information related to other component logics as alien information and *hide it*- don't unpack the hidden information until we have reduced a given problem to a sub-problem in the relevant component logic. Hence a good understanding and

[8]This strategy called *embedding* is the one adopted by Finger and Gabbay in [36].

practical methodology for combining logics could provide vital tools in the design of a complex system. Such a methodology can help decompose the problem of designing a complex system into developing components (logics) and combining them.

Combining logics suggests identifying the basic components of a system, define the corresponding formal tools and combine them to produce a *scalable* formalisation of the whole system. This methodology would allow the user to re-use previously defined and understood components (logics) and would guide him on how to combine these in a proficient way. Put in simple terms the problem of combining logics is this: given two logics Λ_1 and Λ_2, how to combine them into a single logic $\Lambda_1 \otimes \Lambda_2$ which extends the expressive power of each one. For example, suppose Λ_1 addresses temporal aspects of agents and Λ_2 addresses epistemic aspects. Their combination should be able to express both temporal and epistemic properties, but also the interaction of these two aspects: evolving knowledge, and knowledge about a changing world. Even if the logics are similar it might be the case that they are presented in different ways; for example Λ_1 may be described by an axiomatisation, while we may have only the semantics of Λ_2. Hence a methodology for combining systems has now become essential for most applications. Recent trends in logic-based agent technology shows that any logical system modelling agents should be a combined system of logics of knowledge, belief, time and modal logics (of actions). But the problem of combining systems can be pretty difficult in practice, not only because the two systems to be combined may be presented in two completely different ways, but also because that even when they are presented in the same way, it is not clear how to combine them. It is therefore necessary to identify the problems involved in combining logics and come up with a uniform working definition of a logic system that allows to define combination of the different logics involved. The problem of combining logics can be defined as follows:[9]

1. Given two logics Λ_1 and Λ_2 in languages $\mathcal{L}_1$ and $\mathcal{L}_2$ can we define a logic Λ for the combined language, $(\Lambda_1 \otimes \Lambda_2)$, which is a conservative extension of each of Λ_1 and Λ_2?
   Define the syntax of $\Lambda_1 \otimes \Lambda_2$ *from the syntax of* Λ_1 *and* Λ_2*.*

2. If Λ_1 and Λ_2 are complete with respect to classes of models (it could be $\mathcal{K}_1$ and $\mathcal{K}_2$ of Kripke models), is it possible to identify natural semantic constructions that would yield a new semantics for Λ $(\Lambda_1 \otimes \Lambda_2)$?
   Define the semantics of $\Lambda_1 \otimes \Lambda_2$ *from the semantics of* Λ_1 *and* Λ_2*.*

[9] We adopt the following symbols (1) $\oplus$ Addition of Axioms (2) $\otimes$ Combination of logics (3) $\circledast$ Fusion of logics (4) $\odot$ Fibring of logics.

3. How many options for Λ are there and how do they relate to each other?

4. Is it possible for Λ_1 and Λ_2 to interact? And if so how?

5. If Λ_1 and Λ_2 are presented as proof systems Π_1 and Π_2 can we find a combinatorial definition for combining them?
   Define the proof theory of $\Lambda_1 \otimes \Lambda_2$ *from* Λ_1 *and* Λ_2.

6. How can we account for transfer of properties from Λ_1 and Λ_2 to Λ?
   Prove transfer of important properties like soundness,completeness, decidability, finite model property etc. of the logics Λ_1 *and* Λ_2 *into* $\Lambda_1 \otimes \Lambda_2$.

7. If Λ is already given then is it possible to decompose Λ into Λ_1 and Λ_2 where Λ_i are projections onto the sub-languages $\mathbf{L}_i$, so that Λ could be reconstructed back as some combination of Λ_1 and Λ_2 with possibly additional interaction axioms?

8. Is it possible to characterise known poly-modal logics?

A few different techniques [35, 77, 46, 47, 129, 128] have been put forward to address some of these problems but it is still a long way to go before we have a comprehensive understanding of the issue at hand. We give importance to two such techniques viz. *fusion* [77] and *fibring* [42] and recast BDI logics in terms of fibring.

2.3.1 Fusion of Modal Logics

The formation of *fusions* or *independent joins*, is the simplest and perhaps most frequently used way of combining logics. This operation is done on the proof theoretic level when two proof theoretic systems are combined. Kracht and Wolter [77] investigate the transfer properties for a combination between two mono-modal normal logics into a particular normal bi-modal logic. Their definition runs as follows: Given a set Φ of propositional variables $\mathfrak{p}_0, \mathfrak{p}_1, \ldots$, and primitive connectives $\wedge, \neg, \Box_1, \Box_2$, let $\mathcal{L}_{\Box_1\Box_2}$, be the language of the bimodal logic built from it. Then $\mathcal{L}_{\Box_1}$ denotes the fragment of $\Box_2$-free formula and $\mathcal{L}_{\Box_2}$ denotes the fragment of $\Box_1$-free formula. We can then say that Λ is a *normal bi-modal logic* if $\Lambda \subseteq \mathcal{L}_{\Box_1\Box_2}$ and Λ satisfies Definition 5 (i.e., closed under axiom **K** and the inference rules of necessitation (for both connectives) and uniform substitution). If Λ is a normal bimodal logic then $\Lambda_{\Box_1} = \Lambda \cap \mathcal{L}_{\Box_1}$ and $\Lambda_{\Box_2} = \Lambda \cap \mathcal{L}_{\Box_2}$ are normal mono-modal logics which means that the projections of a normal bi-modal

logic Λ onto its two constituent languages are normal mono-modal logics. If $\Lambda = \Lambda_{\Box_1} \circledast \Lambda_{\Box_2}$ then Λ is called *independently axiomatisable*. It is also the case that given two mono-modal logics Λ_1 and Λ_2 formulated in languages $\mathcal{L}_1$ and $\mathcal{L}_2$ (but containing the language **PL** of propositional logic) and having disjoint sets of modal operators we can form the *fusion* of $\Lambda_1 \circledast \Lambda_2$, which is the least bi-modal logic Λ in the language $\mathcal{L}_{1,2}$ containing $\Lambda_1 \cup \Lambda_2$ What this means is that if Λ_1 is axiomatised by a set of axioms Ax_1 and Λ_2 is axiomatised by a set of axioms Ax_2, then $\Lambda_1 \circledast \Lambda_2$ is axiomatised by the union $Ax_1 \cup Ax_2$. This means that no axioms containing modal operators from both languages $\mathcal{L}_1$ and $\mathcal{L}_2$ is required to axiomatize the fusion of Λ_1 and Λ_2. The modal operators in $\mathcal{L}_1$ and $\mathcal{L}_2$ remain *independent*, i.e., they do not *interact*. In another sense we can argue that the fusion of two logics Λ_1 and Λ_2 need not be equal to the union of Λ_1 with Λ_2. For instance $\Lambda = (\Lambda_1 \circledast \Lambda_2)$ is closed under substitution and hence if Λ_1 and Λ_2 are normal mono-modal logics with operators $\Box_1, \Box_2$, then $\Box_2(\Box_1 \mathfrak{p}_0 \Rightarrow \mathfrak{p}_1) \Rightarrow (\Box_2 \Box_1 \mathfrak{p}_0) \Rightarrow (\Box_2 \mathfrak{p}_1)$ belongs to $\Lambda_1 \circledast \Lambda_2$ [10].

Definition 16 *If Λ_1, Λ_2 are normal modal logics, the fusion $\Lambda_1 \circledast \Lambda_2$ is the least normal bi-modal logic containing the two.*

In a similar manner one can define the fusion $\Lambda_1 \circledast \Lambda_2 \circledast \ldots \circledast \Lambda_n$ of n logics for any natural number $n \geq 2$ as the formation of fusions is an associative binary operation on logics. Hence the definitions and theorems will be related to multi-modal version rather than to the bi-modal one. In addition to the syntactic interpretation done so far with regard to fusions, we can also define a semantic counterpart for logics which are Kripke-complete. Consider two classes $\mathfrak{F}_1$ and $\mathfrak{F}_2$ of m- and n-frames that are closed under disjoint union and are isomorphic copies. The fusion $\mathfrak{F}_1 \circledast \mathfrak{F}_2$ of $\mathfrak{F}_1$ and $\mathfrak{F}_2$ is the class of all $n+m$-frames of the form $\langle \mathsf{W}, \mathsf{R}_1, \ldots, \mathsf{R}_m, S_1, \ldots, S_n \rangle$ such that $\langle \mathsf{W}, \mathsf{R}_1, \ldots, \mathsf{R}_m \rangle \in \mathfrak{F}_1$ and $\langle \mathsf{W}, S_1, \ldots, S_m \rangle \in \mathfrak{F}_2$. $\mathfrak{F}_1 \circledast \mathfrak{F}_2$ consists of arbitrary combinations of frames from $\mathfrak{F}_1$ and $\mathfrak{F}_2$ sharing the same set of worlds. The case is that if $\mathfrak{F}_1$ and $\mathfrak{F}_2$ determine logics Λ_1 and Λ_2, then all frames in $\mathfrak{F}_1 \circledast \mathfrak{F}_2$ validate the fusion $\Lambda_1 \circledast \Lambda_2$.

Theorem 13 *[45]* ***(Fusion of logics preserves Kripke completeness)*** *If multi-modal logics Λ_1 and Λ_2 are characterised by the classes of frames $\mathfrak{F}_1$ and $\mathfrak{F}_2$, respectively, and if $\mathfrak{F}_1$ and $\mathfrak{F}_2$ are closed under the formation of disjoint unions and isomorphic copies, then the fusion $\Lambda_1 \circledast \Lambda_2$ of Λ_1 and Λ_2 is characterised by $\mathfrak{F}_1$ and $\mathfrak{F}_2$*

[10] The example is taken from [85].

A number of other properties have been proved for fusion of bi-modal logics by Kracht and Wolter. They give a word of caution saying that though their results could be generalised to logics with arbitrary many modal operators some care should be taken in the case of certain properties (like finite model property). We list some of the important properties

Theorem 14 *[77] (**Properties of Fusion**) If Λ_1 and Λ_2 are two consistent normal modal logics then the following properties hold*

1. The logic $\Lambda_1 \circledast \Lambda_2$ is finitely axiomatisable iff both Λ_1 and Λ_2 are and $\bot \notin \Lambda_1, \Lambda_2$
2. The logic $\Lambda_1 \circledast \Lambda_2$ is complete iff both Λ_1 and Λ_2 are and $\bot \notin \Lambda_1, \Lambda_2$
3. The logic $\Lambda_1 \circledast \Lambda_2$ is compact iff both Λ_1 and Λ_2 are and $\bot \notin \Lambda_1, \Lambda_2$
4. The logic $\Lambda_1 \circledast \Lambda_2$ has finite model property (f.m.p) iff both Λ_1 and Λ_2 have f.m.p and $\bot \notin \Lambda_1, \Lambda_2$
5. If the logics Λ_1 and Λ_2 are complete, then the logic $\Lambda_1 \circledast \Lambda_2$ is decidable iff both Λ_1 and Λ_2 are and $\bot \notin \Lambda_1, \Lambda_2$
6. If the logics Λ_1 and Λ_2 are complete, then the logic $\Lambda_1 \circledast \Lambda_2$ is $Hallde'n-$*complete* [11] iff both Λ_1 and Λ_2 are and $\bot \notin \Lambda_1, \Lambda_2$
7. If the logics Λ_1 and Λ_2 are complete, then the logic $\Lambda_1 \circledast \Lambda_2$ has interpolation[12] iff both Λ_1 and Λ_2 have interpolation and $\bot \notin \Lambda_1, \Lambda_2$.

Some of the results like decidability, interpolation and f.m.p as given above for normal bi-modal logics do hold in the case of multi-modal logics as shown by Gabbay [45]. Gabbay gives the following theorem in the case of f.m.p.

Theorem 15 *[45](**Finite Model Property**) If both Λ_1 and Λ_2 are Kripke complete multi-modal logics having the finite model property, then their fusion $\Lambda_1 \circledast \Lambda_2$ has the finite model property as well.*

[11] A logic Λ is said to be $Hallde'n-complete$ if $\varphi \vee \psi \in \Lambda$ and $var(\phi) \cap var(\psi) = \emptyset$ implies $\varphi \in \Lambda$ or $\psi \in \Lambda$, where $var(\phi)$ and $var(\psi)$ are the sets of propositional variables that appear in ϕ and ψ respectively. The notion of $Hallde'n-completeness$ is related to the concept of *relevance*.

[12] A logic Λ has the interpolation property if whenever $\varphi \Rightarrow \psi \in \Lambda$ then there is a formula ϕ with $var(\phi) \subseteq var(\varphi) \cap var(\psi)$ such that $\varphi \Rightarrow \phi \in \Lambda$ and $\varphi \Rightarrow \phi \in \Lambda$. The formula ϕ is called the interpolant for $\phi \Rightarrow \psi$ in Λ.

Though the results given by Kracht [77] are strong and it is possible to extend the process of fusion to normal n-ary modal logics it does not give any result if we want to express interactions between the respective logics. The earliest papers dealing with transfer issues [35, 36, 77] all deal with combining logical systems without interaction. Negative results in the presence of interaction has been shown in [127, 66, 91, 90, 129]. Hence we need a more general technique which allows us to accommodate the respective interaction axioms.

2.3.2 Fibring/Dovetailing Modal Logics

Fibring is one of the most general approaches for combining logical systems through their semantics and took shape in the works of Gabbay [40, 41, 42]. The methodology allows the user to combine the semantics of the two systems and *weave* the two proof-theories into a combined logic that preserves the basic properties of the components. Though Gabbay [42] shows that fibring could be used as a general methodology for combining logics irrespective of their presentation and type, the most successful case of fibring is in the case of modal logic. The reason for this is that the modality operators allows one to speak directly about the possible worlds where the local fibring is being done. In this thesis we are concerned with modal fibring based on Kripke semantics.

On the other side of the spectrum, much work has been done of late concerning property transfer of logics based on algebraic semantics [134, 17] as well as defining new semantics for the fibred logics [112, 27]. For them, the essential concept of *fibred semantics* is the internal structure of models and not necessarily the notion of world [26]. In our work we do not explicitly give a fibred semantics for BDI logics, but show the necessary conditions the fibring function ought to satisfy in-order to accomodate types of *interaction* axioms. Ongoing research from the algebraic side is also concerned with that of analysing first-order modal logic in terms of the fibring methodology so as to come up with a new semantics as a result of fibring modal logic and first-order logic [113, 115]. Work has also been done relating fibring and para-consistent logics [26]. Our aim in this work is to analyse BDI-logics in terms of fibring based on the work by Gabbay [42] and account for certain transference of properties from the component logics to the fibred one.

As stated earlier the fibring of two logics is obtained by combining their languages, deduction systems and semantics. For this to happen an analysis of the syntax, semantics and proof-theory of the respective logics need to be done. In the case of fibring one can have two views regarding the atomic structure which leads to two notions of fibring called *disjoint fibring*

and *ordinary fibring*. Let $\mathfrak{Q}_1$ and $\mathfrak{Q}_2$ be the sets of atoms from which the languages $\mathcal{L}_1$ and $\mathcal{L}_2$ are built. If $\mathfrak{Q}_1 \cap \mathfrak{Q}_2 = \emptyset$ i.e., if the languages have no atoms in common, then the fibring of $\mathcal{L}_1 \circledcirc \mathcal{L}_2$ is called *disjoint fibring* and if $\mathfrak{Q}_1 = \mathfrak{Q}_2 = \mathfrak{Q}$ then it is called *ordinary fibring*. For our work we prefer *ordinary fibring*. Consider two modal logics Λ_1 and Λ_2 and their corresponding languages $\mathcal{L}_1, \mathcal{L}_2$ sharing the same set $\mathfrak{Q}$ of atomic propositions. Let it also be the case that the logics are presented via classes $\mathcal{K}_1$, $\mathcal{K}_2$ of Kripke models and satisfaction relations $\models_1, \models_2$. As far as fibring is concerned there is no need to assume that the classes are frame classes or normal or anything of that sort. The aim is to define a logic that has the expressivity of both the components. To combine them syntactically is easy as one can consider the formation rules of both languages $\mathcal{L}_1, \mathcal{L}_2$. The problem arises when we try to define the models and satisfaction relations (i.e., when we try to define them semantically). Suppose that $\Box_1$ and $\Box_2$ are two modalities belonging to the $\mathcal{K}_1$, $\mathcal{K}_2$ classes of models as mentioned above. We know how to evaluate $\Box_1\varphi$ in $\mathcal{K}_1$ (i.e. in $\mathfrak{M}^1 = (\mathcal{F}^1, \omega^1, \nu^1)$) and $\Box_2\varphi$ in $\mathcal{K}_2$ (i.e., in $\mathfrak{M}^2 = (\mathcal{F}^2, \omega^2, \nu^2)$)[13] and propositional formulas in both. But the difficulty arises when we try to interpret a formula composed of operators from both logics (for example $\Box_1\Box_2(\varphi)$. We can inductively interpret $\Box_1$ in one of the models for Λ_1 ($\mathfrak{M}^1 = (\mathcal{F}^1, \omega^1, \nu^1)$)[14] and then interpret the rest. In order to interpret the modal operator $\Box_2$ we have to use a model for Λ_2 and in this case it is possible only if we link (*fibre*), via a *fibring function*, the model for Λ_1 with a model for Λ_2 and build a fibred model of the combination. In other words each time we need to evaluate a formula φ of the form $\Box_2\varphi$ in a world in a model of $\mathcal{K}_1$ we associate via the fibring function $\mathfrak{F}$, to the world a model in $\mathcal{K}_2$ where we calculate the truth value of the formula. Formally

$$\omega \models_{\mathfrak{M}^1 \in \mathcal{K}_1} \Box_2\varphi \text{ if and only if } \mathfrak{F}_{\mathfrak{M}^1}(\omega) \models_{\mathfrak{M}^2 \in \mathcal{K}_2} \Box_2\varphi$$

φ holds in ω iff it holds in the model associated to W through the fibring function $\mathfrak{F}$. In more general terms this could be given as

$$\models_\omega \varphi \text{ if and only if } \models^*_{\mathfrak{F}(\omega)} \varphi$$

where $\mathfrak{F}$ is a fibring function that maps a world to a model suitable for interpreting φ and $\models^*$ is the corresponding satisfaction relation (either $\models_1$ for Λ_1 or $\models_2$ for Λ_2). The basic idea of fibring is to perform a model construction while calculating the interpretation of a formula. To demonstrate how the fibring function works consider the following example

[13] By $\mathfrak{M}^1 = (\mathcal{F}^1, \omega^1, \nu^1)$ we denote a model in $\mathcal{K}_1$ and by $\mathfrak{M}^2 = (\mathcal{F}^2, \omega^2, \nu^2)$ a model in $\mathcal{K}_2$.

[14] where $\mathcal{F}^1 = (\mathsf{W}^1, \mathsf{R}^1)$.

Example 1 *Consider two modal logics $K\Box_1$ and $KB\Box_2$ and let $\varphi = \Box_1\Diamond_2\mathfrak{p}_0$ be a formula on a world ω_0 of the fibred semantics. φ belongs to the language $\mathcal{L}_{(1,2)}$ as the outer connective ($\Box_1$) belongs to the language $\mathcal{L}_1$ and the inner connective ($\Diamond_2$) belongs to the language $\mathcal{L}_2$.*

By the standard definition we start evaluating $\Box_1$ of $\Box_1\Diamond_2$ at ω_0. Hence according to the standard definition we have to check whether $\Diamond_2\mathfrak{p}_0$ is true at every ω_1 accessible from ω_0 since from the point of view of $\mathcal{L}_1$ this formula has the form $\Box_1 p$ (where $p = \Diamond_2\mathfrak{p}_0$ is atomic) and $\mathcal{F}^1$ does not recognize $\Diamond_2$. But at ω_1 we cannot interpret the operator $\Diamond_2$, because we are in a model of $K\Box_1$, not of $KB\Box_2$. In order to do this evaluation we need the fibring function $\mathfrak{F}$ which at ω_1 points to a world v_0, a world in a model suitable to interpret formula with $\Box_2$ as the main operator (Figure 2.6). Now all we have to check is whether $\Diamond_2\mathfrak{p}_0$, is true at v_0 in this last model and this can be done in the usual way. Hence the fibred semantics for the combined language $\mathcal{L}_{(1,2)}$ has models of the form $(\mathcal{F}^1, \omega^1, \nu^1, \mathfrak{F}^1)$ where $\mathfrak{F}^1$ is the fibring function which associates a model $\mathfrak{M}^2_\omega$ from $\mathcal{L}_2$ with ω in $\mathcal{L}_1$ i.e. $\mathfrak{F}^1(\omega) = \mathfrak{M}^2_\omega$.

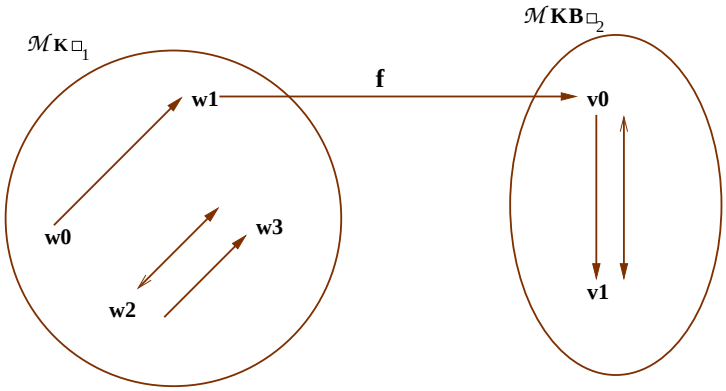


Figure 2.6: An example of Fibring

2.3.3 The language of Modal Fibring/Dovetailing

The basic syntax of fibring includes a family Λ_i, $i \in \mathbf{I}$ (set of labels) of modal logics in the language $\mathcal{L}_i$, consisting of the expressions E_i, built from a set of atomic units $\mathfrak{Q}_i$ and $\mathbb{C}_i$ of constructors of $\mathcal{L}_i$. It is usually the case that $\mathfrak{Q}_1 = \mathfrak{Q}_2 = \mathfrak{Q}$. The expression $E_i = (\mathfrak{q}^i_1, \ldots, \mathfrak{q}^i_n)$ indicates that E_i is built up from the atoms $\mathfrak{q}^i_1, \ldots, \mathfrak{q}^i_n \in \mathfrak{Q}_i$. The semantic construction is the same as in modal logics where Λ_i is characterised by a class $\mathcal{K}_i$ of models $(\mathfrak{M}^i_1, \mathfrak{M}^i_2, \ldots)$

where each $\mathfrak{M}_n^i$ is built from a set W_n^i of basic semantic components called worlds $\omega \in \mathsf{W}_n^i$ and usually has a distinguished point $\mathbf{a}_n^i \in \mathsf{W}_n^i$. This distinguished point called the *actual* world is important in terms of semantic evaluation in the model. It is also assumed that $\mathsf{W}_m^i \cap \mathsf{W}_n^i = \emptyset$ for $m \neq n$. The value of a general expression E_i of Λ_i evaluated at a model $\mathfrak{M}_n^i$ is given as $\nu^i(\mathfrak{M}_n^i, E_i)$ and at a world ω as $\nu^i(\omega, E_i)$ where $\omega \in \mathsf{W}_n^i$. The fibred language $\mathcal{L}_{(a_1,\ldots,a_n)}$ is defined as follows:

- Let $\mathcal{L}_{(i)}$ be $\mathcal{L}_i$, $i \in \mathbf{I}$
- Let $\overline{b}$ be $(j, b_1, \ldots, b_k)$ such that $(j, b_1, \ldots, b_k) \in I$ and $i \neq j$. Let $\mathcal{L}(i) * \overline{b}$ be the family of all expressions of the form $\varphi = E_i(\mathfrak{q}_1/\mathfrak{t}_1 \ldots, \mathfrak{q}_n/\mathfrak{t}_n)$ where $E_i(\mathfrak{q}_1, \ldots, \mathfrak{q}_n) \in \mathcal{L}_i$ and $\mathfrak{t}_1 \ldots, \mathfrak{t}_n$ are in $\mathcal{L}_{\overline{b}}$, and $\mathfrak{q}_x/\mathfrak{t}_x$ indicates the substitution of $\mathfrak{t}_x$ to $\mathfrak{q}_x$ in E_i.
- $\mathcal{L}_{\mathbf{I}} = \bigcup_{\overline{b}} \mathcal{L}_{\overline{b}}$

What this means is that given two modal logics Λ_1 and Λ_2 in languages $\mathcal{L}_1$ and $\mathcal{L}_2$ the language $\mathcal{L}_{1,2}$ of $\Lambda_{1,2}$ contains $\mathcal{L}_1$ and $\mathcal{L}_2$ as well as the expressions obtained from them by substituting atoms with formulas of the other language. So if we consider Λ_i $i \in \mathbf{I}$ to be modal logics in the respective languages $\mathcal{L}_i$ and complete with respect to a class of models $\mathcal{K}_i = \{\mathfrak{M}_1^i, \mathfrak{M}_2^i, \ldots, \}$ then each model $\mathfrak{M}_n^i$ has the form $(\mathsf{W}, \mathsf{R}, \mathbf{a}, \nu)$ where W is the set of possible worlds, $\mathbf{a} \in \mathsf{W}$ is the actual world and $\mathsf{R} \subseteq \mathsf{W}^2$ is the *accessibility relation*. ν being a binary function gives a value $\nu(\omega, \mathfrak{p}_0) \in \{0, 1\}$ for any $\omega \in \mathsf{W}$ and atomic $\mathfrak{p}_0$. $\mathbf{a}$ plays an important role in the semantic evaluation, as satisfaction in the model is defined as satisfaction at $\mathbf{a}$. Gabbay [42] gives the following semantic condition for models

$$\mathsf{W} = \{ \text{ x } | \exists \text{ n } \mathbf{a}\mathsf{R}^n x\}$$

Points not accessible from $\mathbf{a}$ by any power R^n of R do not affect truth values at $\mathbf{a}$ and hence such an assumption is feasible in terms of satisfaction relation. It is also assumed that all sets of possible worlds in any $\mathcal{K}_i$ are all pairwise disjoint and there are infinitely many isomorphic (but disjoint) copies of each model in $\mathcal{K}_i$. Since we are dealing with modal logic fibring, what this means is that each model involved has the form $\mathfrak{M}_n^i = (\mathsf{W}_n^i, \mathsf{R}_n^i, \mathbf{a}_n^i, \nu_n^i)$ and since all W_n^i are pairwise disjoint we can put the semantics together in one big set of possible worlds $\mathsf{W} = \bigcup_{(i,n)}$: In a similar manner the relations R_n^i can all be unified under a single relation. We don't expand on this point here as we give the entire formalism while defining the fibred structure. To make the notations simple we represent a model as $\mathfrak{M} = (\mathsf{W}^{\mathfrak{M}}, \mathsf{R}^{\mathfrak{M}}, \mathbf{a}^{\mathfrak{M}}, \nu^{\mathfrak{M}})$

and write $\mathfrak{M} \in \mathcal{K}_i$ when the model $\mathfrak{M}$ is in the semantics $\mathcal{K}_i$. What this means is that if $\mathfrak{M} \neq \mathfrak{N}$ then $\mathsf{W}^{\mathfrak{M}} \cap \mathsf{W}^{\mathfrak{N}} = \emptyset$ and it is possible to identify a model by its actual world by saying that $\mathfrak{M} = \mathfrak{N}$ iff $\mathbf{a}^{\mathfrak{M}} = \mathbf{a}^{\mathfrak{N}}$. The final definition we have to give in this section is with regard to the fibring function $\mathfrak{F}$. The assumptions to be made in this case are as follows. For two logics Λ_1 and Λ_2 in languages $\mathcal{L}_1$ and $\mathcal{L}_2$ it is required that the values $\nu^2(\mathfrak{M}^2_n, E^2)$ for models $\mathfrak{M}^2_n$ of Λ_2 and E^2 of $\mathcal{L}_2$ are acceptable values for the function $\nu^1(\omega, \mathfrak{p}_0)$, for $\omega \in \mathsf{W}^1$ and $\mathfrak{p}_0$ atomic. Hence for $\mathcal{L}_{(1,2)}$ the fibred models will have the form $\mathfrak{M}^1_n, \mathfrak{F}^{(1,2)}_n$, where $\mathfrak{F}^{(1,2)}_n$ is a fibring function on W^1_n assigning for each $\omega \in \mathsf{W}^1_n$ a model $\mathfrak{M}^2_t \in \mathcal{K}_2$ and the basic evaluation clause for $\omega \in \mathsf{W}^1$ is given as $\nu^{1,2}(\omega, E^2) = \nu^2(\mathfrak{M}^2_\omega, E^2)$. Since for each ω, $\mathfrak{F}^{(1,2)}(\omega)$ is a model $\mathfrak{M}^2_\omega$, an index n_ω is given to the fibred modal so that it is uniquely represented ($\mathfrak{M}^2_\omega = \mathfrak{M}^2_{n_\omega}$). The fibring function $\mathfrak{F}^{(1,2)}_n$ either gives the numerical value n_ω or some actual world $\mathbf{a} \in \mathsf{W}^2_{n_\omega}$ so that any $\mathbf{a}_\omega \in \mathsf{W}^2$ will characterise a model $\mathfrak{M}^2_{n_\omega}$. In fact it is possible to allow $\mathfrak{F}^{(1,2)}(\omega)$ to be $\mathbf{a}^2_{n_\omega} \in \mathsf{W}^2$, as the models always have a distinguished world called the actual world $\mathbf{a}^i_n \in \mathsf{W}^i_n$. Another assumption made with regard to the fibring function is that $\mathfrak{F}^{(1,2)}$ is a one-to-one function from W^1 into W^2 satisfying the condition

$$\omega \neq \omega' \text{ where } \omega, \omega' \in \mathsf{W}^1 \Rightarrow n_\omega \neq n_{\omega'}$$

What this means is that if we have $\omega \in \mathsf{W}^1_k$ and $\omega' \in \mathsf{W}^1_m$ and $k < m$ then the fibred models $\mathfrak{F}^{(1,2)}(\omega) = \mathbf{a}^2_{n_\omega}$ and $\mathfrak{F}^{(1,2)}(\omega') = \mathbf{a}^2_{n_{\omega'}}$ satisfy $n_\omega < n_{\omega'}$. To summarise, if $\omega \in \mathsf{W}^1$ then $\mathfrak{F}(\omega) = \mathbf{a}^2_{n_\omega} \in \mathsf{W}^2$ and if $\omega \in \mathsf{W}^2$ then $\mathfrak{F}(\omega) = \mathbf{a}^1_{n_\omega} \in \mathsf{W}^1$. Basically, given a model $\mathfrak{M}$ and ω as a possible world in $\mathfrak{M}$ then $\mathfrak{F}(\mathfrak{M}, \omega)$ is a Kripke model of the other semantics. A formula φ holds in ω, $\omega \models_{\mathfrak{M}} \varphi$, if and only if $\mathfrak{F}(\mathfrak{M}, \omega) \models \varphi$, i.e. if and only if $\mathbf{a}^{\mathfrak{F}(\mathfrak{M},\omega)} \models \varphi$ where φ is a formula with main connective not in the logic of the model $\mathfrak{M}$ and $\mathbf{a}^{\mathfrak{F}(\mathfrak{M},\omega)}$ is the actual world of the model $\mathfrak{F}(\mathfrak{M}, \omega)$. The properties given so far with regard to modal fibring can be given in the form of a definition as follows

Definition 17 *Let $\Lambda_i, i \in I$, be a family of modal logics with $\Box_i \Lambda \in I$ respectively. Let $\mathcal{K}_i$ be a class of models for which Λ_i is complete. A fibred model for the logic $\Lambda^{\mathfrak{F}}_i$ is a structure $(\mathsf{W}, \mathsf{W}_{i,i\in I}, \mathsf{W}_a, \mathsf{R}, \omega_0, \nu, \mathfrak{F})$ where*

- $\mathsf{W} = \bigcup_{\mathfrak{M} \in \bigcup_i \mathcal{K}_i} \mathsf{W}^{\mathfrak{M}}$;
- $\mathsf{W}_{i,i\in I} = \{a^{\mathfrak{M}} \mid \mathfrak{M} \in \mathcal{K}_i\}$;
- $\mathsf{W}_a = \bigcup_i \mathsf{W}_i$;

- $\mathsf{R} = \bigcup_{\mathfrak{M} \in \cup_i \mathcal{K}_i} \mathsf{R}^{\mathfrak{M}}$;
- $\omega_0 \in \mathsf{W}_a$ *is the actual world;*
- $\nu(\omega, \mathfrak{q}) = \nu^{\mathfrak{M}}(\omega, \mathfrak{q})$, *for the unique* $\mathfrak{M}$ *such that* $\omega \in \mathsf{W}^{\mathfrak{M}}$;
- $\mathfrak{F} : I \times \mathsf{W} \mapsto \mathsf{W}_i$, *is the fibring function.*

The fibring function $\mathfrak{F}$ *is a function giving for each* i *and each* $\omega \in \mathsf{W}$ *another point (actual world) in* W_i *as follows:*

$$\mathfrak{F}_i(\omega) = \begin{cases} \omega & \text{if } \omega \in \mathsf{W}^{\mathfrak{M}} \text{ and } \mathfrak{M} \in \mathcal{K}_i \\ \text{a value in } \mathsf{W}_i, & \text{otherwise} \end{cases}$$

such that if $\omega \neq \omega'$ *then* $\mathfrak{F}_i(\omega) \neq \mathfrak{F}_i(\omega')$.

Satisfaction is defined as follows with the usual truth tables for boolean connectives:

$$\begin{array}{ll} \omega \models \mathfrak{q} & \text{iff } \nu(\omega, \mathfrak{q}) = 1, \text{ where } \mathfrak{q} \text{ is an atom} \\ \omega \models \Box_i \varphi & \text{iff } \begin{cases} \omega \in \mathfrak{M}^i \text{ and } \forall \omega' (\omega \mathsf{R} \omega' \rightarrow \omega' \models \mathfrak{q}_1) \text{ or} \\ \omega \in \mathfrak{M}, \text{ and } \mathfrak{M} \notin \mathcal{K}_i \text{ and } \mathfrak{F}_i(\omega) \models \Box_i \varphi \end{cases} \end{array}$$

We say the model satisfies φ *iff* $\omega_0 \vDash \varphi$.

In a similar manner we can define

$$\omega \models \Diamond_i \varphi \quad \text{iff} \quad \begin{cases} \omega \in \mathfrak{M}^i \text{ and } \exists \omega' (\omega \mathsf{R} \omega' \wedge \omega' \models \mathfrak{q}_1) \text{ or} \\ \omega \in \mathfrak{M}, \text{ and } \mathfrak{M} \notin \mathcal{K}_i \text{ and } \mathfrak{F}_i(\omega) \models \Diamond_i \varphi \end{cases}$$

We say the model satisfies φ iff $\omega_0 \models \varphi$. In the case of validity we say that φ is valid if for all models of $\Lambda_i^{\mathfrak{F}}$ with actual world $\omega_0 \in \mathsf{W}$ we have $(\omega_0, \mathbf{a}^{\omega_0}) \models \varphi$.

Theorem 16 *[42] (**Completeness theorem for the fibred logic** $\Lambda_I^{\mathfrak{F}}$)*
Let $\Lambda_i, i \in I$ *be modal logics in the respective language* $\mathcal{L}_i$ *with classes of structures* $\mathcal{K}_i$ *and set of theorems* $\mathfrak{T}_i$*(i.e.,* $\mathfrak{T}_i = \{\varphi \text{ of } \mathcal{L}_i \mid \varphi \text{ is valid in all } \mathcal{K}_i \text{ models } \}$*)*
Let $\Lambda_i^{\mathfrak{F}}$ *be defined as follows:*

1. $\Lambda_i \subseteq \Lambda_i^{\mathfrak{F}}$ *for any* $i \in I$
2. ***Modal Fibring Rule*** *If* $i \neq j$, *and* $\varphi = (\bigwedge_{k=1}^{n} \Box_i \psi_k \Rightarrow \bigvee_{k=1}^{m} \Box_i \phi_k) \in \Lambda_I^{\mathfrak{F}}$, *then for all* d, $\Box_j^d \varphi \in \Lambda_I^{\mathfrak{F}}$;

3. $\Lambda_i^{\mathfrak{F}}$ *is the smallest set closed under 1, 2, modus ponens and uniform substitution.*

Gabbay [42] proves the following properties:

Theorem 17 ***(Properties of the fibred modal logic)***

1. $\Lambda_I^{\mathfrak{F}}$ *is the set of valid formulae of every model in* $\mathfrak{M}_{\Lambda\mathfrak{F}_i}$.

2. *If all the* $\Lambda_i, i \in I$, *satisfy finite model property, then so does* $\Lambda_i^{\mathfrak{F}}$.

3. *If all the* Λ_i *are finitely axiomatisable, so is* $\Lambda_i^{\mathfrak{F}}$.

So far we were concerned with the general notion of fibring and didn't mention anything about *dovetailing* which is a special case of *ordinary* fibring. As mentioned earlier ordinary fibring happens when the languages $\mathcal{L}_1$ and $\mathcal{L}_2$ share the same set of atoms $\mathfrak{Q}$. We noted in the section above that $\mathfrak{F}$ can be viewed as a function giving for each $\omega \in \mathsf{W}^1 \cup \mathsf{W}^2$ an element $\mathfrak{F}(\omega) \in \mathsf{W}^1 \cup \mathsf{W}^2$ such that if $\omega \in \mathsf{W}^i$ then $\mathfrak{F}(\omega) \in \mathsf{W}^j, i \neq j$. Now if we compare the values $\nu^i(\omega, \mathfrak{q})$ and $\nu^i(\mathfrak{F}(\omega), \mathfrak{q})$ for atom $\mathfrak{q}$ they need not be identical. But it is possible that the fibring function $\mathfrak{F}$ could account for a valuation like

$$\nu^i(\omega, \mathfrak{q}) = \nu^i(\mathfrak{F}(\omega), \mathfrak{q})$$

for each $\omega \in \omega^i, \mathfrak{F}(\omega) \in \mathsf{W}^j$ and each $\mathfrak{q} \in \mathfrak{Q}$. Such instances of fibring is called *dovetailing* i.e. instances in which the fibred model at ω has the world ω itself as its actual world.

Definition 18 *Let* $\Lambda_i, i \in I$ *be modal logics with* $\mathcal{K}_i$, *the class of models for* Λ_i. *Let* $\Lambda_I^{\mathfrak{D}}$ *(the dovetailing combination of* $\Lambda_i, i \in I$*) be defined semantically through the class of all (dovetailed) models of the form* $(\mathsf{W}, \mathsf{R}, \mathbf{a}, \nu)$, *where* W *is a set of worlds,* $\mathbf{a} \in \mathsf{W}$, ν *is a value assignment, and for each* $i \in I, \mathsf{R}(i) \subseteq \mathsf{W} \times \mathsf{W}$. *We require that for each* i, $(\mathsf{W}, \mathsf{R}(i), \mathbf{a}, \nu)$ *is a model in* $\mathcal{K}_i$. *We further require the following:*
Let $\omega \in \mathsf{W}$ *be such that there exist* $n_1, \ldots, n_k$ *and* $\omega_1, \ldots, \omega_k$ *such that* $\mathbf{a}\mathsf{R}^{n_1}(\omega_1) \circ \mathsf{R}^{n_2}(\omega_2) \ldots \circ \mathsf{R}^{n_k}(\omega_k)\omega$ *holds;*
The notion of $\omega \models \varphi$ *by induction is defined as follows.*

- $\omega \models \mathfrak{q}$ *if* $\nu(\omega,) = 1$ *for* $\mathfrak{q}$ *atomic.*
- $\omega \models \Box_i\varphi$ *if for all* $\omega' \in \mathsf{W}$, *such that* $\omega\mathsf{R}(i)\omega'$ *we have* $\omega' \vDash \varphi$.
- $\models \varphi$ *iff for all models and actual worlds* $\mathbf{a} \models \varphi$.

Definition 19 *Let $\Lambda_i, i \in I$, be modal logics; let $\Lambda_I^{\mathfrak{D}}$ be defined as follows:*
1. $\Lambda_i \subseteq \Lambda_I^{\mathfrak{D}}$
2. ***M****odal Dovetailing Rule: If $i \neq j$, and*
$\varphi = (\bigwedge_{k=1}^{n} \Box_i \psi_k \wedge \bigwedge_{k=1}^{m} \Diamond_i \neg \phi_k \Rightarrow \bigvee_{k=1}^{m} \mathfrak{q}_k) \in \mathfrak{T}_i^{\mathfrak{D}}$,
then for all $d, \Box_j^d \varphi \in \mathfrak{T}_i^{\mathfrak{D}}$, where $\mathfrak{q}_k$ are atoms (or their negations) and $\mathfrak{q}_1, \ldots, \mathfrak{q}_r$ list all the atoms (or their negations) in any $\psi_k, \phi_k, k = 1, \ldots$
3. $\Lambda_I^{\mathfrak{D}}$ *is the smallest set closed under 1, 2, modus ponens and substitution.*

The following properties hold in the case of dovetailed logics.

Theorem 18 *(**Properties of the dovetailed modal logic**)*

1. *$\Lambda_I^{\mathfrak{D}}$ is the set of valid formulae of every model in $\mathfrak{M}_{\Lambda_I^{\mathfrak{D}}}$.*
2. *If each of the $\Lambda_i, i \in I$ include the K axiom and can be formulated by an Hilbert style system with necessitation rule, then $\Lambda_I^{\mathfrak{D}}$ can be axiomatised by taking the union of the axiomatisations.*
3. *If all the Λ_i satisfy finite model property, then so does $\Lambda_I^{\mathfrak{D}}$.*

It has been shown in [35] that on the semantic level fusion of two normal 1-modal logics corresponds to *dovetailing*. For instance, let Λ_1 and Λ_2 be two mono-modal logics and $\mathcal{F}_1$ and $\mathcal{F}_2$ the respective frames for which they are complete. Then the fusion $\Lambda_1 \circledast \Lambda_2$ is complete with respect to the class of all frames obtained by iterated dovetailing constructions from $\mathcal{F}_1$ and $\mathcal{F}_2$. Gabbay has also shown that in some cases fibring and dovetailing produce the same result. In particular

Theorem 19 *If each of the $\Lambda_i, i \in I$ admits necessitation and satisfies the disjunction property*[15] *then $\Lambda_I^{\mathfrak{F}} = \Lambda_I^{\mathfrak{D}}$.*

2.4 Summary and Discussion

In this chapter we first introduced modal logics (its syntax, semantics) along with the idea of canonical models which helps in proving completeness of a system of modal logic. We then introduced BDI logics and showed that they are normal multimodal logics with an arbitrary set of normal modal operators and specific interaction axioms. It was shown that any BDI system modelling rational agents consists of a *combined* system of logics of belief,

[15] If $\models \Box\varphi \Rightarrow (\Box\psi \vee \Box\phi)$, then $\models \Box\varphi \Rightarrow \Box\psi$ or $\models \Box\varphi \Rightarrow \Box\phi$.

goal and intentions. In order to account for this combined nature we introduced two combining techniques called *fusion* and *fibring* and showed that fibring is more suited to combine BDI logics that fusion. Our idea behind adopting such combining techniques is to provide a general methodology for combining the component logics involved in BDI-like systems.

Though results like Theorem 18 makes one think of the chances of fibring being boiled down to fusion, in principle fibring is more powerful than fusion because of the possibility of adding conditions on the fibring function. These conditions could encode interactions between the two classes of models that are being combined and therefore could represent interaction axioms between the two logics. Moreover fibring is not restricted to normal modal logics. Gabbay shows how to extend it to some non-standard logic systems. Adaptations of the fibring technique can apparently be applied to many other cases as well. It is possible to fibre a logic with a fuzzy logic, fibre a first order logic with itself, combine a logic with its meta level etc.

The main thrust of fibring as far as BDI logics is concerned is given in the following statement [85]:

> It is sometimes possible to recognize some existing combined systems as fibrings or dovetailings but difficulties arise when the combination is not a simple fusion, but an interaction between the components is present. By investigating some class of fibring functions it will be possible to give a basic class of interactions and prove properties about these.

As BDI logics are nothing but combinations of basic normal logics it is possible to analyse them in terms of fibring/dovetailing and get some general results about them. In the next chapter we show that the BDI-system is a dovetailed logic. In order to accommodate the interaction axioms involved in BDI we give conditions on the fibring function. Though fusion offers transfer theorems for properties like completeness, decidability etc. for both single modal and multi-modal settings its inability to account for interaction axioms makes it less suited for BDI-like logics. Hence we prefer fibring over fusion.

Chapter 3

Fibring Semantics for BDI Logics

Logic takes care of itself; all we have to do is to look and see how it does it.

Ludwig Wittgenstein

In this chapter we examine BDI logics in the context of Gabbay's [42] *fibring* semantics. We show that *dovetailing* (a special form of fibring) can be adopted as a semantic methodology to characterise BDI logics. We develop a set of interaction axioms that can capture static as well as dynamic aspects of the mental states in BDI systems, using Catach's *incestual* schema $G^{a,b,c,d}$ [19]. Further we exemplify the constraints required on fibring function to capture the semantics of interactions among modalities. The advantages of having a fibred approach is discussed in the final section.

3.1 Background and motivations

In the previous chapter we showed that any BDI system modelling rational agents consists of a combined system of logics of belief, goal and intentions. Moreover, many more logics are studied for modeling BDI-like agents wherein a combination of knowledge, time and modal logic of actions is considered [123, 103, 111]. These combined systems was presented and motivated by different authors for different reasons and different applications. However, the general methodology for combining the different logics involved has been mainly neglected. We believe that investigating a general methodology for combining the component logics involved in BDI-like systems is an important research issue. This would result in a better understanding of the formal groundings of complex rational agent architectures

and enable the designer to elegantly and easily incorporate new features of rational agency within her framework. Moreover the proposed general methodology should permit a modular treatment of the modal components, whereby, each component is analysed and developed on its own, with the most appropriate methodology for it, and is then reused in the combination. Furthermore each module has its own features but the framework remains unchanged among the combined systems. Finally the combined system should offer preservation of some important logical properties of its elements.

In this chapter we investigate one such method, viz. *fibring* [42], and use it to reconstruct the logical account of BDI in terms of *dovetailing* together with the multi-modal semantics of Catach [19]. In doing so we identify a set of interaction axioms for BDI, based on the incestual schema $G^{a,b,c,d}$, which covers many of the existing BDI axioms and also make possible the generation of a large class of new ones. Further, we identify conditions under which completeness transfers from the component logics (Λ_1 and Λ_2) to their *fibred/dovetailed* composition ($\Lambda_{1,2}^{\mathfrak{F}}/\Lambda_{1,2}^{\mathfrak{D}}$), with the help of canonical model structures. We also show completeness preservation in the case of interaction axiom of the form $\Box_1\alpha \Rightarrow \Box_2\alpha$ ($\Lambda_{1,2}^{\mathfrak{F},\mathfrak{D}} \oplus \Box_1\alpha \Rightarrow \Box_2\alpha$). Our study differs from that of other combining techniques like *fusion* in terms of the interaction axiom. For instance, normal bimodal and polymodal logics without any interaction axioms are well-studied as *fusions* of normal monomodal logics in [128] and property transfer for such logics has also been dealt with [77]. For a slightly different account on fusions of logics one can refer [86]. Moreover *fusions* of normal modal logics without interaction axioms is the same as *dovetailing*. But difficulty arises with the extra interaction axiom. Then we need a more general concept like *fibring*. Our study starts with the assumption that the combination of two complete logics need not be complete when we add interaction axioms [76]. We want to identify conditions under which completeness can be preserved when interaction axioms are included.

3.2 The Problem

As noted in the previous chapter, the main advantage of using Multi-Modal Logics in BDI is their ability to express complex modalities, that can capture the inter-relationships existing among the different mental attitudes. This can be achieved by either composing modal operators of different types, or by using formal operations over modalities. For instance the idea that an agent's goal is always supported by its belief is captured by the following

BDI axioms:

$$\text{GOAL}^{\mathbf{KD}}(\varphi) \Rightarrow \text{BEL}^{\mathbf{KD45}}(\varphi) \tag{3.1}$$

$$\text{GOAL}^{\mathbf{KD}}(\varphi) \Rightarrow \text{BEL}^{\mathbf{KD45}}(\text{GOAL}^{\mathbf{KD}}(\varphi)) \tag{3.2}$$

Accordingly, the semantic conditions for (1) and (2) are:

$$\text{if } (\omega', \omega'') \in \text{BEL then } (\omega', \omega'') \in \text{GOAL} \tag{3.3}$$

$$\text{if } (\omega', \omega'') \in \text{BEL and } (\omega'', \omega''') \in \text{GOAL then } (\omega', \omega''') \in \text{GOAL} \tag{3.4}$$

Condition (3.3) captures inclusion (containment) of a binary relation for beliefs in the relation for goals, whereas (3.4) captures the combined transitivity on two binary relations R_1 and R_2. The problem here is that the axiom systems for GOAL and BEL is a combination of other axiom systems and hence they are different. They can be considered as two different languages $\mathcal{L}_1$ and $\mathcal{L}_2$ with $\Box_1$ (BEL) and $\Box_2$ (GOAL) built up from the respective sets $\mathfrak{Q}_1$ and $\mathfrak{Q}_2$ of atoms and supported by the logics $\mathbf{KD45}(\Lambda_1)$ and $\mathbf{KD}(\Lambda_2)$. Hence we are dealing with two different systems S_1 and S_2 characterized, respectively, by the class of Kripke models $\mathcal{K}_1$ and $\mathcal{K}_2$ and this fact should be taken into consideration while defining semantic conditions for interaction axioms like those given above. For instance, we know how to evaluate $\Box_1\varphi$ (BEL(φ)) in $\mathcal{K}_1$ ($KD45$) and $\Box_2\varphi$ (GOAL(φ)) in $\mathcal{K}_2$ (KD). We need a method for evaluating $\Box_1$ (resp. $\Box_2$) with respect to $\mathcal{K}_2$ (resp. $\mathcal{K}_1$).

The problem in its general form is how to construct a multi-modal logic containing several unary modalities, each coming with its own system of specific axioms. The fibring technique as described in the previous chapter allows one to combine systems through their semantics. The fibring function can evaluate (give a yes/no) answer with respect to a modality in S_2, being in S_1 and vice versa. Each time we have to evaluate a formula φ of the form $\Box_2\psi$ in a world in a model of $\mathcal{K}_1$ we associate, via the fibring function $\mathfrak{F}$, to the world a model in $\mathcal{K}_2$ where we calculate the truth value of the formula. Formally

$$\omega \models_{\mathfrak{M}\in\mathcal{K}_1} \Box_2\psi \text{ iff } \mathfrak{F}_2(\omega) \models_{\mathfrak{M}'\in\mathcal{K}_2} \Box_2\psi$$

ψ holds in ω iff it holds in the model associated to w through the fibring function $\mathfrak{F}$. Moreover ψ could be a mixed wff consisting of operators from Λ_1 and Λ_2 (for instance ψ can be $\Diamond_1\Box_2 q$). This is possible because the axiom systems of BEL and GOAL itself are combined systems. Then we have to say that the wff ψ belongs to the language $\mathcal{L}_{(1,2)}$. The existing BDI semantics fails to give adequate explanation for such formulas. The

problem becomes even more complex when we allow the system to vary in time. Then we have to combine the BDI system with a suitable temporal logic. The fibring/dovetailing technique provides a general methodology for such combinations as shown in the next section.

It is also important to note that since each mental operator (BEL, GOAL) itself is a combination of different axiom systems, the underlying multi-modal language ($\mathcal{L}_{\mathbf{BDI}}$) should be such that we should be able to develop each single operator on its own within its own semantics so that in the later stage the operators and models can be glued together through fibring/dovetailing to form a comprehensive system. The multi-modal language should also be able to express multiple concepts like rational agents, actions, plans etc. The problem with the existing BDI-Language is that each time we want to incorporate a new concept we have to redefine the system and come up with characterization results. For instance, if we want to capture the notion of actions, plans, programs etc. in BDI, we need to come up with specific axiom systems for each of them and then show that they are characterized within the system. What we need is a set of interaction axioms that can generate a range of multi-modal systems for which there is a general characterization theorem so that we could avoid the need for showing it each time a new system is considered. To this end we adopt the class of interaction axioms $G^{a,b,c,d}$ of Catach [19] that can account for a range of multi-modal systems.

3.3 Fibring of BDI Logics

In this section we show how fibring/dovetailing could be adopted as a general semantic methodology to combine BDI logics. Two theorems stating relationships between dovetailing and BDI logics and dovetailing and fibring are shown. It is shown that the existing BDI logic is a dovetailed system.

Fibring two semantics

Let $\mathbf{I}$ be a set of labels representing intentional states (belief, goal, intention) and $\Lambda_i, i \in \mathbf{I}$ be modal logics whose respective modalities are $\Box_i, i \in \mathbf{I}$.

Definition 20 *[42] A fibred model is a structure* $(\mathsf{W}, \mathsf{S}, \mathsf{R}, \mathbf{a}, \nu, \tau, \mathbf{F})$ *where*

- W *is a set of possible worlds;*
- S *is a function giving for each* ω *a set of possible worlds,* $\mathsf{S}^\omega \subseteq \mathsf{W}$*;*
- R *is a function giving for each* ω*, a relation* $\mathsf{R}^\omega \subseteq \mathsf{S}^\omega \times \mathsf{S}^\omega$*;*

- $\mathbf{a}$ *is a function giving the actual world* $\mathbf{a}^\omega$ *of the model labelled by* ω*;*
- ν *is an assignment function* $\nu^\omega(\mathfrak{q}_0) \subseteq \mathsf{S}^\omega$*, for each atomic* $\mathfrak{q}_0$*;*
- τ *is the semantical identifying function* $\tau : \mathsf{W} \to \boldsymbol{I}$*.* $\tau(\omega) = i$ *means that the model* $(\mathsf{S}^w, \mathsf{R}^\omega, \mathbf{a}^\omega, \nu^\omega)$ *is a model in* $\mathcal{K}_i$*, we use* W_i *to denote the set of worlds of type* i*;*
- $\mathbf{F}$*, is the set of fibring functions* $\mathfrak{F} : \boldsymbol{I} \times \mathsf{W} \mapsto \mathsf{W}$*. A fibring function* $\mathfrak{F}$*is a function giving for each* i *and each* $\omega \in \mathsf{W}$ *another point (actual world) in* W *as follows:*

$$\mathfrak{F}_i(\omega) = \begin{cases} \omega & \text{if } \omega \in \mathsf{S}^{\mathfrak{M}} \text{ and } \mathfrak{M} \in \mathcal{K}_i \\ \text{a value in } \mathsf{W}_i, & \text{otherwise} \end{cases}$$

such that if $\omega \neq \omega'$ *then* $\mathfrak{F}_i(\omega) \neq \mathfrak{F}_i(\omega')$*. It should be noted that fibring happens when* $\tau(\omega) \neq i$*. Satisfaction is defined as follows with the usual truth tables for boolean connectives:*

$$t \models \mathfrak{q}_0 \text{ iff } \nu(\omega, \mathfrak{q}_0) = 1, \text{ where } \mathfrak{q}_0 \text{ is an atom}$$
$$\omega \models \Box_i\varphi \text{ iff } \begin{cases} \omega \in \mathfrak{M} \text{ and } \mathfrak{M} \in \mathcal{K}_i \text{ and } \forall\omega'(\omega \mathsf{R} \omega' \to \omega' \models \varphi), \text{or} \\ \omega \in \mathfrak{M}, \text{ and } \mathfrak{M} \notin \mathcal{K}_i \text{ and } \forall \mathfrak{F} \in \mathbf{F}, \mathfrak{F}_i(\omega) \models \Box_i\varphi. \end{cases}$$

We say the model satisfies φ *iff* $\omega_0 \models \varphi$*.*

A fibred model for $\Lambda_{\mathbf{I}}^{\mathfrak{F}}$ can be generated from fibring the semantics for the modal logics $\Lambda_i, i \in \mathbf{I}$. The detailed construction runs as follows: Let $\mathcal{K}_i$ be a class of models $\{\mathfrak{M}_1^i, \mathfrak{M}_2^i, \ldots\}$ for which Λ_i is complete. Each model $\mathfrak{M}_n^i$ has the form $(\mathsf{S}, \mathsf{R}, \mathbf{a}, \nu)$ where, as given in the second chapter, S is the set of possible worlds, $\mathbf{a} \in \mathsf{S}$ is the actual world and $\mathsf{R} \subseteq \mathsf{S} \times \mathsf{S}$ is the accessibility relation. ν is the assignment function, a binary function, giving a value $\nu(\omega, \mathfrak{p}_0) \in \{0, 1\}$ for any $\omega \in \mathsf{S}$ and atomic $\mathfrak{p}_0$. The actual world $\mathbf{a}$ plays a role in the semantic evaluation in the model, in so far as satisfaction in the model is defined as satisfaction at $\mathbf{a}$. We can assume that the models satisfy the condition $S = \{\omega \mid \exists n \ \mathbf{a} \ \mathsf{R}^n\omega\}$. This assumption does not affect satisfaction in models because points not accessible from $\mathbf{a}$ by any power R^n of R do not affect truth values at $\mathbf{a}$. Moreover we assume that all sets of possible worlds in any $\mathcal{K}_i$ are all pairwise disjoint, and that there are infinitely many isomorphic (but disjoint) copies of each model in $\mathcal{K}_i$. We use the notation $\mathfrak{M}$ for a model and present it as $\mathfrak{M} = (\mathsf{S}^{\mathfrak{M}}, \mathsf{R}^{\mathfrak{M}}, \mathbf{a}^{\mathfrak{M}}, \nu^{\mathfrak{M}})$ and write $\mathfrak{M} \in \mathcal{K}_i$, when the model $\mathfrak{M}$ is in the semantics $\mathcal{K}_i$. Thus our assumption boils down to $\mathfrak{M} \neq \mathfrak{N} \Rightarrow \mathsf{S}^{\mathfrak{M}} \cap \mathsf{S}^{\mathfrak{N}} = \varnothing$. In fact a model can be identified by its actual world, i.e., $\mathfrak{M} = \mathfrak{N}$ iff $a^{\mathfrak{M}} = a^{\mathfrak{N}}$. Then the fibred semantics can be given as follows:

- $\mathsf{W} = \bigcup_{\mathfrak{M} \in \cup_i \mathcal{K}_i} \mathsf{S}^{\mathfrak{M}}$;
- $\mathsf{R} = \bigcup_{\mathfrak{M} \in \cup_i \mathcal{K}_i} \mathsf{R}^{\mathfrak{M}}$;
- $\nu(\omega, \mathfrak{q}_0) = \nu^{\mathfrak{M}}(\omega, \mathfrak{q}_0)$, for the unique $\mathfrak{M}$ such that $\omega \in \mathsf{S}^{\mathfrak{M}}$;
- $\mathbf{a}^{\omega} = \mathbf{a}^{\mathfrak{M}}$ for the unique $\mathfrak{M}$ such that $\omega \in \mathsf{S}^{\mathfrak{M}}$.

Dovetailing

As we pointed out in the second chapter, dovetailing is a special case of fibring in the sense that a dovetailed model is a fibred model that must agree with the current world on the values of atoms. For instance, in the previous section we saw that the function $\mathfrak{F}$ can be viewed as functions giving for each $\omega \in \mathsf{S}^1 \cup \mathsf{S}^2$, an element $\mathfrak{F}(\omega) \in \mathsf{S}^1 \cup \mathsf{S}^2$ such that if $\omega \in \mathsf{S}^i$ then $\mathfrak{F}(\omega) \in \mathsf{S}^j, i \neq j$. If $\mathcal{L}_1$ and $\mathcal{L}_2$ share the same set of atoms $\mathfrak{Q}$ then we can compare the values $\nu(\omega, \mathfrak{q}_0)$ and $\nu(\mathfrak{F}(t), \mathfrak{q}_0)$ for an atom $\mathfrak{q}_0$ which need not be identical. If we require from the fibring functions that for each $\omega \in \mathsf{S}^i, \mathfrak{F}_j(t) \in \mathsf{S}^j$ and each $\mathfrak{q}_0 \in \mathfrak{Q}$ we want

$$\nu(\omega, \mathfrak{q}_0) = \nu(\mathfrak{F}_i(\omega), \mathfrak{q}_0),$$

then this fibring case is referred to as *dovetailing*. This means that the actual world of the model fibred at ω, $\mathfrak{F}_i(\omega)$, can be identified with ω. The set of fibring functions $\mathbf{F}$ is no longer needed, since we identified ω with $\mathfrak{F}_i(\omega)$, for every fibring function $\mathfrak{F}$.

Definition 21 *[42] Let Λ_i be modal logics, where $\mathcal{K}_i$ is the class of models for Λ_i. Let $\Lambda_I^{\mathfrak{D}}$ (the dovetailing combination of $\Lambda_i, i \in \boldsymbol{I}$) be defined semantically through the class of all (dovetailed) models of the form $(\mathsf{W}, \mathsf{R}, a, \nu)$, where W is a set of worlds, $a \in \mathsf{W}$, ν is an assignment and for each $i \in \boldsymbol{I}, \mathsf{R}(i) \subseteq \mathsf{W} \times \mathsf{W}$. We require that for each i, $(\mathsf{W}, \mathsf{R}(i), a, \nu)$ is a model in $\mathcal{K}_i$. It is further required that all $t \in \mathsf{W}$ be such that there exist $n_1, \ldots, n_k$ and $i_1, \ldots, i_k$ such that $a\mathsf{R}^{n_1}(i_1) \circ \mathsf{R}^{n_2}(i_2) \cdots \circ \mathsf{R}^{n_k}(i_k)t$ holds. The satisfaction condition $\omega \models \varphi$, is defined by induction as*

- $\omega \models \mathfrak{q}_0$ *if* $\omega \in \nu(\mathfrak{q}_0)$ *for* $\mathfrak{q}_0$ *atomic;*
- $\omega \models \Box_i \varphi$ *if for all* $\omega' \in \mathsf{W}$, *such that* $\omega \mathsf{R}(i) \omega'$ *we have* $\omega' \models \varphi$;
- $\models \varphi$ *iff for all models and actual worlds* $a \models \varphi$.

Two theorems are given below, the proof of which can be found in [42].

Theorem 20 (Dovetailing and Normal Modal Logic) *Assume Λ_i, $i \in \boldsymbol{I}$ all are extensions of $\boldsymbol{K}$ formulated using traditional Hilbert axioms and the rule of necessitation, then $\Lambda_I^{\mathfrak{D}}$ can be axiomatized by taking the union of the axioms and the rules of necessitation for each modality $\Box_i$ of each Λ_i*

Theorem 21 (Fibring = Dovetailing) *If Λ_i, $i \in \boldsymbol{I}$ admit necessitation and satisfy the disjunction property, then $\Lambda_I^{\mathfrak{F}} = \Lambda_I^{\mathfrak{D}}$.*

It is immediate to see that BDI logic without interaction axioms is nothing else but normal multi-modal logics —combinations of normal modal logics (e.g., the basic BDI logic proposed in [107] is the combination of a **KD45** modality for BEL, **KD** for GOAL and **KD** for INT)— hence, according to Theorem 1, dovetailing provides a general methodology for generating BDI-like systems.

3.4 Semantics for Mental States

In the previous section we have seen how to provide a general semantics for BDI logics in the background of fibring/dovetailing without taking into account any interaction of modalities. However, mental states are very often connected to each other, for example the interaction axioms like (3.1) and (3.2); thus what is needed is a methodology to capture them. In this section we use Catach approach [19] to extend dovetailing in order to develop a general semantics that covers both the basic modalities and their interactions. Catach proposed a class of normal multimodal logics that is determined by the *interaction axiom* $G^{a,b,c,d}$, called a, b, c, d-incestuality axiom. It includes most of the modal and multimodal systems given in chapter 2. The idea is to *label* modal operators with complex parameters using an operator of *composition* and an operator of *union*.

3.4.1 Catach's Incestual Schema

Let I be a set of atomic labels (denoting the mental states beleif, goal and intention); complex labels can be built from atomic ones using the sequential operator ";" (a binary operator for *composition*), the union operator "$\cup$" and the neutral element "λ" (w.r.t the composition). If i is an atomic label and φ a well-formed formula, then the expression $[i]\alpha$ corresponds to the modal formula $\Box_i\varphi$ (or the belief formula BEL(φ)), and $\langle i\rangle\alpha$ to $\Diamond_i\alpha$. Furthermore we assume that $[\lambda] = \langle\lambda\rangle$. The transformation of complex labels into modalities is governed by the following rules:

$$[\lambda]\varphi \Leftrightarrow \varphi \tag{3.5}$$

$$[a;b]\varphi \Leftrightarrow [a][b]\varphi \tag{3.6}$$

$$[a \cup b]\varphi \Leftrightarrow [a]\varphi \wedge [b]\varphi \tag{3.7}$$

According to the above conditions we can identify, for example, the formula

$$\Box_1\Box_2 A \wedge \Box_3 A \wedge A$$

with the expression $[(1;2) \cup 3 \cup \lambda]$. Let us consider now the expression

$$\langle a\rangle[b]\alpha \Rightarrow [c]\langle d\rangle\alpha$$

known as the a, b, c, d-*incestuality* axiom, (we will use $G^{a,b,c,d}$ to refer to it), where a, b, c, and d are strings of modalities obtained from composition and union as shown above. As it is remarked in [19], the fact that a, b, c, and d of an incestual schema may be arbitrary expressions built from atomic labels using the composition and union operators, makes axiom $G^{a,b,c,d}$ very general. In particular it covers the axiom

$$G^{k,l,m,n} : \Diamond^k\Box^l\varphi \Rightarrow \Box^n\Diamond^n\varphi$$

where $k, l, m, n \geq 0$, as is given in section 2.1.4.. Therefore it can be used to generate, among others, the well known **D**, **T**, **B**, **4** and **5** and their multimodal versions. For example, when $a = b = \lambda$ and $c = d = 1$ we obtain the symmetry axiom B for $\Box_i$. See Table 3.1 for a list of such axioms. It is then immediate to see that the above axiom schema covers many existing systems of multi-modal logic, including the BDI system and make the generation of a large class of new ones possible.

Example 2 *Let φ be a formula and* BEL*,* GOAL*,* INT*,* CAP*,* OPP *and* RES *be the modal operators for the mental constructs; then the following are instances of $G^{a,b,c,d}$.*

C1 $\text{GOAL}(\varphi) \Rightarrow \text{BEL}(\varphi)$ (Inclusion)

C2 $\text{INT}(\varphi) \Rightarrow (\text{GOAL}(\varphi) \Rightarrow \text{BEL}(\varphi))$ (Relative Inclusion)

C3 $\text{RES}(\varphi) \Leftrightarrow \text{CAP}(\varphi) \wedge \text{OPP}(\varphi)$ (Union)

The axioms C2 and C3 are possible additions to the existing BDI axioms. The above axioms (as well as others) can be used to represent various concepts such as rational agents, programs, actions etc. For instance C2 captures the fact that an agent's intention to achieve φ is supported by having a goal towards φ and this goal is based on its belief of φ. Similarly

Axiom Name	Axiom Schema	Incestual Schema
B	$\varphi \Rightarrow [i]\langle i\rangle\varphi$	$G^{\lambda,\lambda,i,i}$
T	$[i]\varphi \Rightarrow \varphi$	$G^{\lambda,i,\lambda,\lambda}$
D	$[i]\varphi \Rightarrow \langle i\rangle\varphi$	$G^{\lambda,i,\lambda,i}$
4	$[i]\varphi \Rightarrow [i][i]\varphi$	$G^{\lambda,i,(i;i),\lambda}$
5	$\langle i\rangle\varphi \Rightarrow [i]\langle i\rangle\varphi$	$G^{i,\lambda,i,i}$
Inclusion	$[i]\varphi \Rightarrow [i']\varphi$	$G^{\lambda,i,i',\lambda}$
Semi-commutativity	$[i][i']\varphi \Rightarrow [i'][i]\varphi$	$G^{\lambda,(i;i'),(i';i),\lambda}$
Relative Inclusion	$[i]\varphi \Rightarrow ([i']\varphi \Rightarrow [i'']\varphi)$	$G^{\lambda,(i\cup i'),(i''),\lambda}$
Semi-adjunction	$\varphi \Rightarrow [i]\langle i'\rangle\varphi$	$G^{\lambda,\lambda,i,i'}$
Common seriality	$[i]\varphi \Rightarrow \langle i'\rangle\varphi$	$G^{\lambda,i,\lambda,i'}$
Union	$[i]\varphi \Rightarrow [i']\varphi \wedge [i'']\varphi$ $[i']\varphi \wedge [i'']\varphi \Rightarrow [i]\varphi$	$G^{\lambda,i,(i'\cup i''),\lambda}$ $G^{\lambda,(i'\cup i''),i,\lambda}$
Composition	$[i]\varphi \Rightarrow [i'][i'']\varphi$ $[i'][i'']\varphi \Rightarrow [i]\varphi$	$G^{\lambda,i,(i';i''),\lambda}$ $G^{\lambda,(i';i''),i,\lambda}$

Table 3.1: Some well known axiom schemas included by the incestual axioms

C3 is related to the *ability* and *opportunity* of an agent to perform an action α so as to obtain a particular *result*[1].

As far as dovetailed models are concerned it is possible to define a mapping ρ between labels and the accessibility relations of dovetailed models.

Definition 22 *Let a and b be labels, i an atomic label, and $(\mathsf{W}, \mathsf{R}(i), \mathbf{a}, \nu)$ a dovetailed model. Then*

$$\rho(i) = \mathsf{R}(i); \quad \rho(\lambda) = \Delta; \quad \rho(a;b) = \rho(a)|\rho(b); \quad \rho(a \cup b) = \rho(a) \cup \rho(b);$$

where the operators $\cup$ (union) and $|$ (composition) are defined for binary relations, and Δ is the diagonal relation over W (i.e., $\rho(\lambda) = \Delta$, where $\Delta = \{(\omega, \omega) \mid \omega \in \mathsf{W}\}$, the identity *relation).*

Definition 23 *Let a, b, c, and d be labels. A dovetailed model $\mathfrak{D} = (\mathsf{W}, \mathsf{R}(i), \mathbf{a}, \nu)$ enjoys the a, b, c, d-*incestuality *property iff the following condition holds for $\mathfrak{D}$.*

$$\rho(a)^{-1}|\rho(c) \subseteq \rho(b)|\rho(d)^{-1}. \tag{3.8}$$

where $\rho(i)^{-1}$ is the *converse* relation of $\rho(i)$. The incestuality condition is shown in Figure 3.1 and can be reformulated as follows:

[1]We talk more about the *ability*, *opportunity* and *result* constructs in Chapter 3.

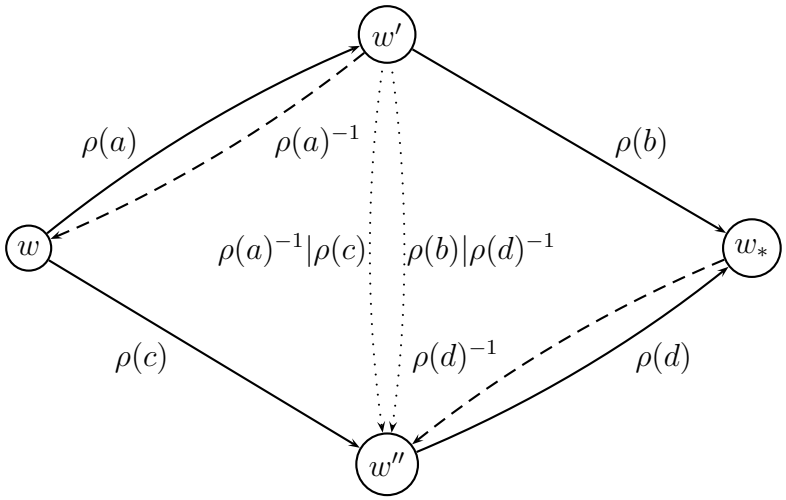


Figure 3.1: a, b, c, d incestuality property

If $(\omega, \omega') \in \rho(a)$ and $(\omega, \omega'') \in \rho(c)$ then there exists ω^* such that $(\omega', \omega^*) \in \rho(b)$ and $(\omega'', \omega^*) \in \rho(d)$.

Theorem 22 *[19] Let $\mathcal{L}_{BDI}$ be a normal multi-modal system built from a finite set of axioms $\boldsymbol{G}^{a,b,c,d}$. Then $\mathcal{L}_{BDI}$ is determined by the class of dovetailed models satisfying the a,b,c,d-incestuality properties.*

Catach originally proved the above theorem for what he calls multi-frames. Trivially multi-frames correspond to dovetailed models. In particular this result provides the characterization of a wide class of interaction axioms such as the relationships among mental attitudes of rational agents in terms of dovetailing.

Although the class of incestual modal logics includes a wide range of multimodal systems as was shown above, it is worth noting that no set of inclusion properties of the form (3.8) can characterise the *indution axiom* that define both the *iteration* operator "$*$" of dynamic logic [65] and the *common knowledge* operator "C" [51, 64]. In [19] the following pair of a,b-induction axioms are given:

$$[b]\varphi \Rightarrow ([a]\varphi \wedge [a][b]\varphi) \tag{3.9}$$

$$[b](\varphi \Rightarrow [a]\varphi) \Rightarrow ([a]\varphi \Rightarrow [b]\varphi) \tag{3.10}$$

By taking $b = a^*$ we get the axioms of dynamic logic and taking $a = 1 \cup 2 \cup \ldots \cup n$ we get the axiom for the common knowledge operator. Axiom

3.9 is an incestual axiom corresponding to

$$\langle\lambda\rangle[b]\varphi \Rightarrow [a \cup a; b]\langle\lambda\rangle\varphi$$

but 3.10 is not as shown in [19].

3.5 Conditions on the Fibring Function

Section 3.3 establishes how BDI-like systems (without interaction axioms) can be reconstructed using dovetailing and section 3.4 introduces a general axiom schema through which we can generate a range of BDI-like interaction axioms. In this section we demonstrate with the help of an example what conditions would be required on the fibring functions in order to cope with the a, b, c, d-incestuality schema. As noted earlier we assume that the combination of two complete logics need not be complete when we include interaction axioms. We want to identify conditions under which completeness can be preserved. But before identifying the specific conditions on the fibring functions we need to introduce certain notions and constructions related to completeness preservation in terms of canonical models and canonical logics.

In the canonical model construction, a *world* is a maximal consistent sets of wff. Thus for any normal propositional modal system S, its canonical model $\langle \mathsf{W}, \mathsf{R}, \nu\rangle$ is defined as follows:

- $\mathsf{W} = \{\omega : \omega$ is a maximal S-consistent set of wff $\}$;
- For any pair of worlds ω and any $\omega' \in \mathsf{W}$, $\omega\mathsf{R}\omega'$ iff $\{\varphi : \Box\varphi \in \omega\} \subseteq \omega'$;
- For any variable $\mathfrak{p}_0$ and any $\omega \in \mathsf{W}$, $\nu(\mathfrak{p}_0, \omega) = 1$ iff $\mathfrak{p}_0 \in \omega$.

But in the case of a fibred model the above construction needs to be modified accordingly as follows: Let $\Lambda_i, i \in \mathbf{I}$ be monomodal normal logic with languages $\mathcal{L}_i$. Let Σ_Λ be the set of all Λ-maximal consistent sets of formula. Given a set S of formulas, $\Lambda^{\Box_i}(S) = \{\varphi : \Box_i\varphi \in S\}$ and $\Lambda^{\mathcal{L}_i}(S) = \{\varphi : \varphi = \Box_i\psi$ or $\varphi = \Diamond_i\psi, \varphi \in S\}$. The canonical model for $\Lambda_I^{\mathfrak{ff}}$, $C_I^{\mathfrak{ff}}$, is the structure $\langle \mathsf{W}, S, \mathsf{R}, \mathbf{F}, a, \tau, \nu\rangle$, where

- $\mathsf{W} = \Sigma_\Lambda \times \mathbf{I}$.
- S is a function $\mathsf{W} \mapsto \wp\mathsf{W}$ such that $S^\omega = \{(x, i) \in \mathsf{W} : \tau(\omega) = i\}$. In other words the set of worlds of the same type as ω.
- $\mathsf{R}^\omega \subseteq S^\omega \times S^\omega$ such that $x\mathsf{R}^\omega y$ iff $\Lambda^{\Box_{\tau(\omega)}}(x) \subseteq y$.

- $\mathbf{F}$ is the set of functions $\mathfrak{F} : \mathbf{I} \times \mathsf{W} \mapsto \mathsf{W}$ (fibring functions) such that

$$\mathfrak{F}_i, (x, j) = \begin{cases} (x, j) & i = j \\ (x, i) & x = a^\omega \\ (y, i) & \text{otherwise} \end{cases}$$

  where $\Lambda^{\mathcal{L}_i(x)} \subseteq y$, and if $x \neq y$, then $\mathfrak{F}_i(x) \neq \mathfrak{F}_j(y)$.

- $a^\omega = \omega$.
- $\tau(x, i) = i$
- $\nu(\mathfrak{p}_0, \omega) = 1$ iff $\mathfrak{p}_0 \in \omega$, for $\mathfrak{p}_0$ atomic.

Lemma 8 *For every formula φ and every world ω in the canonical model*

$$\nu(\omega, \varphi) = 1 \text{ iff } \varphi \in \omega.$$

Proof. The proof is by induction on the complexity of φ. The only difference with the proof of the monomodal case is when $\varphi = \Box_i\psi$ and $\tau(\omega) \neq i$. If $\nu(\omega, \Box_i\psi) = 1$, then for every $\mathfrak{F} \in \mathbf{F}$ $\nu(\mathfrak{F}_i(\omega), \Box_i\psi) = 1$, and we can apply the standard construction for modalities and we obtain that $\Box_i\psi \in \mathfrak{F}_i(\omega)$. Let us now suppose that $\Box_i\psi$ is not in ω. Since ω is maximal $\neg\Box_i\psi \in \omega$; thus $\Diamond_i\neg\psi \in \omega$. $\Lambda^{\mathcal{L}_i} \subseteq \mathfrak{F}_i(\omega)$, hence $\Diamond_i\neg\psi \in \mathfrak{F}_i(\omega)$, from which we derive a contradiction. Thus $\Box_i\psi \in \omega$. The other direction is similar. □

As an immediate consequence of the Lemma we have the following theorem.

Theorem 23 $\Lambda_I^{\mathfrak{F}} \vdash \varphi$ *iff* $C_I^{\mathfrak{F}} \models \varphi$.

Definition 24 *Let $\mathcal{F}_\Lambda$ be the frame of the canonical model for Λ. Λ is canonical iff for every valuation ν, $(\mathcal{F}_\Lambda, \nu)$ is a model for Λ.*

Clearly the above definition is equivalent to the usual condition for a modal logic to be canonical (i.e., that the frame of the canonical model is a frame for Λ). However the fibring construction inherits the valuation functions from the underlying models, and we can obtain different logics imposing conditions on the fibring functions based on the assignments of the variables. The fibred frame for $\mathcal{L}_{1,2}$ is obtained in the same way as the fibred model, replacing the occurrences of models with frames.

Lemma 9 *Let $\mathfrak{M}_\zeta^{\mathfrak{F}} = (\mathsf{W}, S, \mathsf{R}, \mathbf{F}, a, \tau, \nu)$ be the canonical model for $\Lambda_I^{\mathfrak{F}}$. Then for each $\omega \in \mathsf{W}$ $(S^\omega, \mathsf{R}^\omega, \nu^\omega)$ is the canonical model for $\tau(\omega)$.*

Proof. By inspection on the construction of the canonical model for $\Lambda_I^{\mathfrak{F}}$. □

From the above Lemma we obtain:

Theorem 24 *Let $\Lambda_i, i \in \boldsymbol{I}$ be canonical monomodal logics. Then $\Lambda_I^{\mathfrak{F}}$ is canonical.*

For instance the inclusion axiom $\Box_1\varphi \Rightarrow \Box_2\varphi$ is characterized by the dovetailed models where $\mathsf{R}_2 \subseteq \mathsf{R}_1$. However, such a constraint would be meaningless for fibred models where each modality has its own set of possible worlds. So, what is the corresponding condition on fibred models? As we have already seen a fibring function is defined as

$$\mathfrak{F} : \mathbf{I} \times \mathsf{W} \to \mathsf{W}$$

where $\mathbf{I}$ is the set of modalities involved and W is a set of possible worlds. It is worth noting that given a world we can identify the model it belongs to, and that there is a bijection $\mathfrak{M}$ between the actual worlds and their models. So to deal with the inclusion axiom the following constraint must be satisfied:

$$\forall \omega \in \mathsf{W} \; \forall \; \mathfrak{F} \in \mathbf{F} : \mathfrak{M}(\mathfrak{F}_2(w)) \sqsubseteq_N \mathfrak{M}(\mathfrak{F}_1(\omega)) \tag{3.11}$$

where $\sqsubseteq_N$ is the inclusion morphism thus defined:

Definition 25 *Let $\mathfrak{M}_1$ and $\mathfrak{M}_2$ be two models. Then $\mathfrak{M}_2 \sqsubseteq_N \mathfrak{M}_1$ iff there is a morphism $\mathbf{w} : \mathsf{W}_2 \mapsto \mathsf{W}_1$, such that*

- *for each atom $\mathfrak{p}_0$, $\nu_2(\omega, \mathfrak{p}_0) = \nu_1(\mathbf{w}(\omega), \mathfrak{p}_0)$;*
- *if $x\mathsf{R}_2 y$ then $\mathbf{w}(x)\mathsf{R}_1\mathbf{w}(y)$.*

The constraint on the fibring functions to support the *inclusion axiom*, is in alliance with the incestuality axiom $G^{a,b,c,d}$ as stated in the previous section, that is, $\mathsf{R}_2 = \rho(c)$ and $\mathsf{R}_1 = \rho(b)$. The incestuality axiom can be characterised by giving appropriate conditions that identify the (fibred) models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ involved in the inclusion morphism.

It is now possible to provide a characterisation of the fibring/dovetailing of normal modal logics with the incestuality axiom (i.e. $\Diamond_a\Box_b\varphi \Rightarrow \Box_c\Diamond_d\varphi$). But before giving the characterisation theorem we have to take care of some constructions with respect to the labels a, b, c, d of $G^{a,b,c,d}$.

We have to define new cross-modality relations among the possible worlds in a fibred model. Firstly we explain the relation R^i where $i = a, b, c, d$ and this is done at two levels. The first level deals with the construction of labels (atomic as well as complex) without taking the operations of union and composition into consideration. This is given as follows:

1. When $i = a$ and a is atomic, $\Re^a(\omega) = \{\omega' : \tau(\omega) = a \text{ and } \omega \mathsf{R}^\omega \omega'\}$
2. When $i = a$ and $\tau(\omega) \neq a$, $\mathfrak{F}_a(\omega) \mathsf{R}^{\mathfrak{F}_a(\omega)} \omega'$ where $\mathfrak{F}_a(\omega)$ is the particular fibring function, belonging to the set of fibring functions $\mathbf{F}$, associating a model from ω' with the actual world ω.

From this definition and keeping in line with the definition of canonical models as given in the previous section $\omega \mathsf{R}^a \omega'$ iff $\{\varphi : [a]\varphi \in \omega\} \subseteq \omega'$. In a similar manner, at the second level, complex labels in terms of union and compostion could be given as follows

- $\Re^{b;c}(\omega) = \{\omega' : \exists \omega'' : \omega \Re^b \omega'' \wedge \omega'' \Re^c \omega'\}$
- $\Re^{b \cup c}(\omega) = \{\omega' : \omega \Re^b \omega'\} \cup \{\omega' : \omega \Re^c \omega'\}$

From this we can arrive at the following definition

Definition 26 *Let a and b be labels, i an atomic label, and $\langle \mathsf{W}, \mathsf{S}, \mathsf{R}, \mathbf{F}, \mathbf{a}, \tau, \nu \rangle$ a fibred model. Then*

$$\rho(i) = \Re^i; \quad \rho(\lambda) = \Delta; \quad \rho(a;b) = \rho(a)|\rho(b); \quad \rho(a \cup b) = \rho(a) \cup \rho(b);$$

where the operators $\cup$ (union) and $|$ (composition) are defined for binary relations, and Δ is the diagonal relation over W

Definition 27 *Let $\mathfrak{M} = \langle \mathsf{W}, \mathsf{S}, \mathsf{R}, \mathbf{F}, \ Ba, \tau, \nu \rangle$ be a fibred model, a be a label, and $\omega \in \mathsf{W}$. With $\mathfrak{M}(\omega) \upharpoonright \Re^a$ we denote the fibred model $\langle \mathsf{W}', \mathsf{S}, \mathsf{R}, \mathbf{F}, \mathbf{a}, \tau, \nu \rangle$, where $\mathsf{W}' = \{w' \in \mathsf{W} : \omega \Re^a w'\}$.*

Theorem 25 *Let $\Lambda_a, \Lambda_b, \Lambda_c, \Lambda_d$ be canonical normal modal logics and $\Lambda_{abcd} = \Lambda_a \circledcirc \Lambda_b \circledcirc \Lambda_c \circledcirc \Lambda_d$. Then $\Lambda = \Lambda_{abcd} \oplus \Diamond_a \Box_b \varphi \Rightarrow \Box_c \Diamond_d \varphi$ is characterised by the class of fibred models satisfying*

$$\forall \omega \in \mathsf{W} \ , \ \forall \mathfrak{F} \in \mathbf{F} \ , \ \mathfrak{M}^{ac}(\omega) \ \sqsubseteq_N \mathfrak{M}^{bd}(\omega)$$

where

- $\mathfrak{M}^{ac}(\omega) = \mathfrak{M}(\omega) \upharpoonright \rho(a)^{-1} \mid \rho(c)$
- $\mathfrak{M}^{bd}(\omega) = \mathfrak{M}(\omega) \upharpoonright \rho(b) \mid \rho(d)^{-1}$

Proof. For the proof we have to note that, thanks to the fact that Λ_i ($i \in \{a, b, c, d\}$) are canonical, for any pair of world ω and ω', the maximal consistent sets associated with them are the same, i.e., $S^\omega = S^{\omega'}$, they are the set of all the maximal consistent sets. Thus no matter of the fibring

function we choose, we have that the structure of the models obtained from the $\mathfrak{F}_i$'s are the same. Therefore we can use the identity over Σ_Λ as the morphism $\mathbf{g}$ in the inclusion morphism. Moreover by the definition of canonical models the relation $\omega_1 \Re^a \omega_2$ can be given as

$$\omega_1 \Re^a \omega_2 \text{ iff } \{\varphi : [a]\varphi\} \subseteq \omega_2 \tag{3.12}$$

It should be kept in mind that a world ω in the canonical fibred model is a pair, (Σ, t), where Σ is a maximal Λ-consistent set of formulas and t is a label for a modality (check our construction of canonical models). Also we have to be careful about $\Re^a$ as a could be a complex label and hence we need steps 1. and 2. as mentioned above. Accordingly if we examine the fibring of (Σ, t) with respect to the label a, i.e., $\mathfrak{F}_a(\Sigma, t)$, we have two cases:

1. $\tau(\Sigma, t) = a \qquad (t = a)$
2. $\tau(\Sigma, t) \neq a \qquad (t \neq a)$

For (1)

$$(\Sigma, a)\Re^a(\Sigma', a) \text{ iff } \{\varphi : \Box_a\varphi \in \Sigma\} \subseteq \Sigma' \tag{3.13}$$

For (2) we have two subcases

1. if $\mathbf{a}^{(\Sigma,t)} = (\Sigma, t)$ then $\mathfrak{F}_a(\Sigma, t) = (\Sigma, a)$ and then

$$(\Sigma, a)\Re^a(\Sigma', a) \text{ iff } \{\varphi : \Box_a\varphi \in \Sigma\} \subseteq \Sigma' \tag{3.14}$$

2. if $\mathbf{a}^{(\Sigma,t)} \neq (\Sigma, t)$, where $\mathbf{a}$ is a function giving the actual world $\mathbf{a}^\omega$ of the model labelled by ω, then $\mathfrak{F}_a(\Sigma, t) = (\Sigma^*, a)$ such that

$$\Sigma^{\mathcal{L}_a} \subseteq \Sigma^* \text{ and } (\Sigma^*, a)\Re^a(\Sigma', a) \text{ iff } \{\varphi : \Box_a\varphi \in \Sigma^*\} \subseteq \Sigma'$$

and then

$$(\Sigma, t)\Re^a(\Sigma', a) \text{ iff } \{\varphi : \Box_a\varphi \in \Sigma^*\} \subseteq \Sigma' \tag{3.15}$$

In both cases one can find that

$$(\Sigma, t)\Re^a(\Sigma', a) \text{ iff } \{\varphi : \Box_a\varphi \in \Sigma\} \subseteq \Sigma'. \tag{3.16}$$

Hence by the above construction we obtain

$$\omega\Re^a\omega_1 \text{ iff } \{\varphi : \Box_a\varphi \in \omega\} \subseteq \omega_1 \tag{3.17}$$

or equivalently

$$\omega\Re^a\omega_1 \text{ iff } \{\Diamond_a\varphi : \varphi \in \omega_1\} \subseteq \omega \tag{3.18}$$

Now according to the definition of inclusion morphism we have to show that if

$$(\omega_1, \omega_2) \in \rho(a)^{-1} \mid \rho(c)$$

then

$$(\mathbf{g}(\omega_1), \mathbf{g}(\omega_2)) \in \rho(b) \mid \rho(d)^{-1}$$

where $\mathbf{g}$ is the morphism of the inclusion morphism. Now

$$\omega_1, \omega_2 \in \rho(a)^{-1} \mid \rho(c) \text{ iff } \exists z : (\omega_1, z) \in \rho(a)^{-1} \wedge (z, \omega_2 \in \rho(c)).$$

But this corresponds to

$$\exists z, z\Re^a\omega_1 \wedge z\Re^c\omega_2 \text{ (if } z = \omega \text{ then } \omega\Re^a\omega_1 \text{ and } \omega\Re^c\omega_2). \tag{3.19}$$

On the other hand

$$(\mathbf{g}(\omega_1), \mathbf{g}(\omega_2)) \in \rho(b) \mid \rho(d)^{-1} \text{ iff } \exists t : (\mathbf{g}(\omega_1^b, t)) \in \rho(b) \wedge (t, \mathbf{g}(\omega_2)) \in \rho(d)^{-1}$$

and therefore

$$\exists t : \mathbf{g}(\omega_1)\Re^b t \wedge \mathbf{g}(\omega_2)\Re^d t \text{ (if } t = \omega^* \text{ then } \omega_1\Re^b\omega^* \text{ and } \omega_2\Re^d\omega^*) \tag{3.20}$$

Since $\mathbf{g}$ is the identity over Σ_{max} (3.19) and (3.20) imply that for every maximal Λ-consistent sets of sentences ω, ω_1 and ω_2

$$\begin{array}{l}\text{if } \{\varphi : \Box_a\varphi \in \omega\} \subseteq \omega_1 \text{ and } \{\varphi : \Box_c\varphi \in \omega\} \subseteq \omega_2 \text{ then} \\ \{\varphi : \Box_b\varphi \in \omega_1\} \cup \{\varphi : \Box_d\varphi \in \omega_2\} \text{ is } \Lambda\text{-consistent}\end{array}$$

For suppose that this not the case. Then for some $\Box_b\varphi_1, \ldots, \Box_b\varphi_n \in \omega_1$ and some $\Box_d\psi_1, \ldots, \Box_d\psi_m \in \omega_2$ we have

$$\vdash_\Lambda \varphi_1 \wedge \cdots \wedge \varphi_n \Rightarrow \neg(\psi_1 \wedge \cdots \wedge \psi_m)$$

and by the rule $\vdash \alpha \Rightarrow \beta / \vdash \Diamond\alpha \Rightarrow \Diamond\beta$ we can get

$$\vdash_\Lambda \Diamond_d(\varphi_1 \wedge \cdots \wedge \varphi_i) \Rightarrow \Diamond_d\neg(\psi_1 \wedge \cdots \wedge \psi_j) \tag{3.21}$$

and thereby, by applying the $(\Diamond\Box)$-interchange axiom[2], we get

$$\vdash_\Lambda \Diamond_d(\varphi_1 \wedge \cdots \wedge \varphi_i) \Rightarrow \neg\Box_d(\psi_1 \wedge \cdots \wedge \psi_j) \tag{3.22}$$

[2]If φ is any wff which contains an unbroken sequence of $\Box$ and/or $\Diamond$ and ψ results from φ by replacing $\Box$ by $\Diamond$ and $\Diamond$ by $\Box$ throughout that sequence and also inserting or deleting a $\neg$ both immediately before and after that sequence, then $\vdash \varphi \Leftrightarrow \psi$ (and hence if $\vdash \varphi$ then $\vdash \psi$).

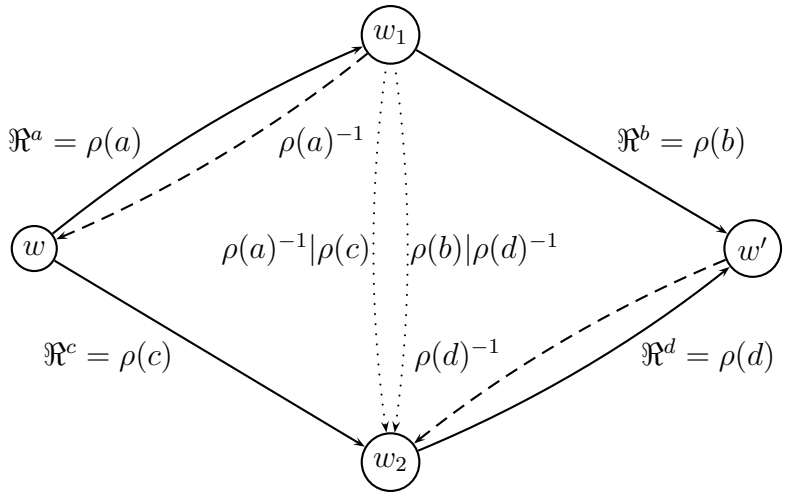


Figure 3.2: Incestuality with respect to cross-modality relations

Since $\Box_b\varphi_1, \ldots, \Box_b\varphi_m \in \omega$, so is $\Box_b(\varphi_1 \wedge \cdots \wedge \varphi_n)$. Hence, by $\omega\Re^a\omega_1$ and by (3.18) we must have

$$\Diamond_a\Box_b(\varphi_1 \wedge \cdots \wedge \varphi_n \in \omega.$$

But ω contains the instance

$$\Diamond_a\Box_b(\varphi_1 \wedge \cdots \wedge \varphi_n) \Rightarrow \Box_c\Diamond_d(\varphi_1 \wedge \cdots \wedge \varphi_n)$$

of $\mathbf{G}^{a,b,c,d}$ and therefore we must have

$$\Box_c\Diamond_d(\varphi_1 \wedge \cdots \wedge \varphi_n) \in \omega.$$

Since $\omega\Re^c\omega_2$, by (3.17), we have

$$\Diamond_d(\varphi_1 \wedge \cdots \wedge \varphi_n) \in \omega_2$$

and, by (3.22), this means that

$$\neg\Box_d(\psi_1 \wedge \cdots \wedge \psi_m) \in \omega_2.$$

which contradicts the assumption that $\Box_d\psi_1, \ldots, \Box_d\psi_m \in \omega_2$, and then $\Box_d(\psi_1 \wedge \cdots \wedge \psi_m) \in \omega_2$. Thus Λ is consistent and completeness is thereby proved. □

Figure 3.2 shows the incestuality condition with respect to cross-modality relations and Figure 3.3 shows how our definition of inclusion morphism fits

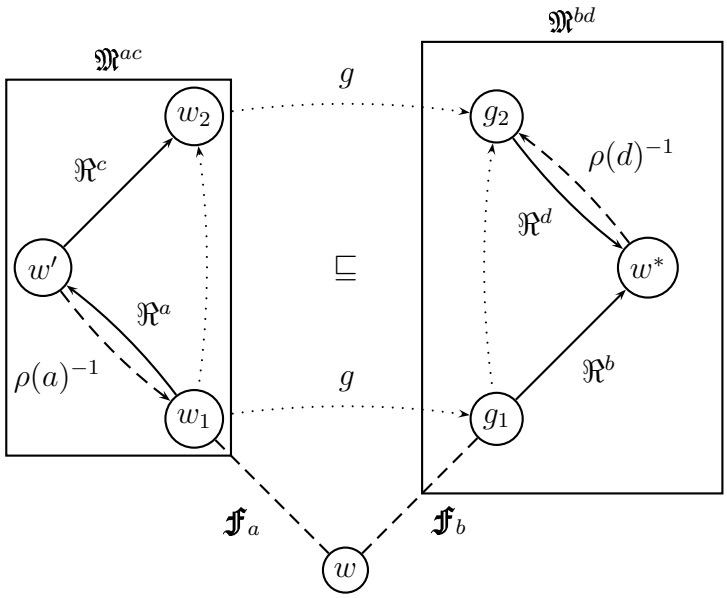


Figure 3.3: Inclusion morphism $\oplus$ fibring $\oplus$ incestuality

in with that of fibring and incestual relation. Now it is time to show how we can account for non-normal operators in such a framework.

It is well known that normal modal operators have certain properties that are occasionally considered undesirable for the common sense notions that they are intended to formalise. For instance the property of *Logical Omniscience* though could hold for the beliefs of an agent is certainly undesirable for the knowledge part. For example to say that an agent knows all the logical consequences of its knowledge $(\Box\varphi \wedge \Box(\varphi \Rightarrow \psi)) \Rightarrow \Box\psi$ is to live in an idealized world. The fibring methodology can be used to combine single modal logics that are not normal. However, in general, simple adjustments are required to deal with classes of non-normal modal logics. In what follows we show the modifications required for quasi-normal modal logics (i.e. modal logics containing K and closed under RM $\vdash \varphi \Rightarrow \psi / \vdash \Box\varphi \Rightarrow \Box\psi$). The first thing we have to consider is that the structure of models appropriate for such a class is $(\mathsf{W}, N, \mathsf{R}, \mathbf{a}, \nu)$ where W and a are as usual, $N \subseteq \mathsf{W}$ (representing the set of normal worlds), $\mathsf{R} \subseteq N \times W$, and we have the following two additional clauses on the valuation function ν:

$$\text{if } \omega \notin N, \nu(\omega, \Box\varphi) = 0; \qquad \text{if } \omega \notin N, \nu(\omega, \Diamond\varphi) = 1.$$

Fibred, dovetailed, and canonical models can be obtained accordingly with

the appropriate trivial modifications (cf. [81]).[3]. We are now ready to give the completeness theorem for the fibring of monotonic modal logics.

Theorem 26 *Let $\Lambda_i, i \in I$ be quasi-normal modal logics classes of structures $\mathcal{K}_i$ and set of theorems $\mathfrak{T}_i$. Let $\mathfrak{T}_I^{\mathfrak{F}}$ be the following set of wffs of $\Lambda_I^{\mathfrak{F}}$.*

1. *$\mathfrak{T}_i \subseteq \mathfrak{T}_I^{\mathfrak{F}}$, for every $i \in I$;*
2. *If $A(x_m) \in \mathfrak{T}_i$ then $A(x_m/\Box_j\varphi_j) \in \mathfrak{T}_I^F$, for any $\Box_i\varphi_m, i \in I$;*
3. ***M**onotonic Modal Fibring Rule: If $\Box_i$ is the modality of Λ_i and $\Box_j$ that of Λ_j, where i, j are arbitrary, with $i \neq j$.*

$$\frac{\bigwedge_{k=1}^{n} \Box_i A_k \Rightarrow \bigvee_{k=1}^{m} \Box_i B_k \in \mathfrak{T}_I^F}{\Box_j^n \bigwedge_{k=1}^{n} \Box_i A_k \Rightarrow \Box_j^n \bigvee_{k=1}^{m} \Box_i B_k \in \mathfrak{T}_I^F} \textit{for all } n;$$

4. *$\mathfrak{T}_I^F$ is the smallest set closed under 1, 2, 3, modus ponens and substitution.*

Then $\mathfrak{T}_I^F$ is the set of all wffs of $\Lambda_I^{\mathfrak{F}}$ valid in all the fibred monotonic structures of $\Lambda_I^{\mathfrak{F}}$.

Proof. The proof is a trivial modification of that of Theorem 3.10 of [42]. □

Intuitively the meaning of the Monotonic Modal Fibring Rule has to do with the substitutions of formulas of one language into a formula of the other language. If the substituted formulas are proof theoretically related we want to propagate this relation to the other language. Moreover there are formal similarities between the Monotonic Fibred Rule and RM. Consider an implication of the form $A \Rightarrow B$ where A and B are built from atoms of the form $\Box_i C$. There our special RM says that if $\vdash A \Rightarrow B$ then we can derive $\vdash \Box_j A \Rightarrow \Box_j B$ for any modality $\Box_j$ other than $\Box_i$.

A similar theorem can be proved for the dovetailing of quasi-normal modal logics with the appropriate modifications on the Dovetail Modal Rule given by Gabbay [42].

At this stage we have to revise our definition of inclusion morphism.

Definition 28 *Let $\mathfrak{M}_1$ and $\mathfrak{M}_2$ be two quasi-normal models. $\mathfrak{M}_1 \sqsubseteq_M \mathfrak{M}_2$ iff*

1. *$\mathfrak{M}_1 \sqsubseteq_N \mathfrak{M}_2$; and*

[3]For normal modal logics $N = W$, thus any normal modal logic is also quasi-normal

2. if $\omega \notin N_1$ *then* $\mathbf{w}(\omega) \notin N_2$.

Theorem 27 *Let* Λ_1 *and* Λ_2 *be the logic obtained by the canonical quasi-normal modal logic fibring/dovetailing of* Λ_1 *and* Λ_2. *Then* $\Lambda_{1,2}^{M} \oplus \Box_1\alpha \Rightarrow \Box_2\alpha$ *is characterized by the class of fibred/dovetailed models satisfying*

$$\forall w \in \mathsf{W} \forall \mathfrak{F} \in \mathbf{F} : M(\mathfrak{F}_2(w)) \sqsubseteq_M M(\mathfrak{F}_1(w)).$$

Proof. The proof is analogous to that of Theorem 25. □

The main consequence of the above theorem is that it shows how to extend the full power of fibring to non-normal modal logics with interaction axioms, including combinations of a range of modalities required to model complex BDI systems.

Corollary 2 *Let* Λ_1 *and* Λ_2 *be the logic obtained by the canonical* {*quasi-*}*normal modal logic fibring/dovetailing of* L_1 *and* L_2. *Then* $L_{1,2}^{M} \oplus \Box_1\alpha \Rightarrow \Box_2\alpha$ *is canonical.*

3.6 Summary and Discussion

We have investigated the relationships between BDI logics and Gabbay's fibring semantics. In particular we have shown how to reconstruct the logical account of BDI in terms of dovetailing and Catach's approach. Roughly fibring (dovetailing) examines and develops each single operator on its own, within its own semantics, and then the operators and models are glued together to form a comprehensive system.

The proposed methodology provides a general framework for BDI in so far as it is not restricted to particular operators and, at the same time, it offers easy tools to study properties (e.g., soundness, completeness, ...) of the resulting systems. The resulting multimodal system is not homogeneous in the sense that it can accommodate interaction axioms of two types. (1) All interactions involved between different mental attitudes of the same agent ($\mathrm{GOAL}_i^{\mathbf{KD}}(\varphi) \Rightarrow \mathrm{BEL}_i^{\mathbf{KD45}}(\varphi)$) and (2) all interactions involved between the same mental attitude of different agents ($\mathrm{GOAL}_i^{\mathbf{KD}} \Rightarrow \mathrm{GOAL}_j^{\mathbf{KD}}$). Such a classification is absent in BDI as proposed by Rao [107]. Transfer results, even if limited to very small cases of interaction axioms, would be of extreme importance to BDI theorists as this will help shift their attention from the single case analysis to the general problem of combination of mental states. The problem of interaction axioms between logic fragments in a combined logic might become more central among the BDI theorists. Moreover the

proposed approach is not restricted to normal modal operators —the use of non-normal epistemic operators is one of the common strategies to (partially) obviate the problem of logical omniscience.

As we have seen dovetailing is a particular form of fibring, and the latter offers a more fine grained analysis of the concepts at hand. In other words we can say that fibring is more selective than dovetailing. Indeed, there are formulas which are valid under dovetailing but false under some interpretations using fibring; thus some unwanted consequences could be discarded by fibring. Remember that the condition for dovetailing states that sources and targets of the fibring functions agree on the evaluation of atoms. However, this is not required for fibring, so, we can say that fibring can be thought of as a change of perspective in passing from a model (modal operator) to another model (modal operator). This has some important consequences: for example, the interpretation of a piece of evidence may depend on the mental state used to look at it.

Chapter 4

Extending BDI with Composite Actions

The great end of life is not knowledge but action.

Thoms Henry Huxley

In this chapter we extend the BDI-formalism to include composite actions of the type $\pi_1; \pi_2$ (read as π_1 followed by π_2). We argue that the successful execution of such actions depends on the Result (RES) of its components. Further, based on a recent work in BDI [97], we investigate the close connection between the Result of an action performed by a BDI agent and its *capability* of achieving that result. The capability factor is supported using the RES construct and it is shown how the components of a composite action is supported using these two. Further, we introduce an OPP (opportunity) operator which in alliance with *Result* and *Capability* provides a suitable semantics for practical reasoning in BDI.

4.1 Background and Motivations

The BDI architecture considered so far did not address issues related to the concept of *actions* or *events*. Just as one can define the *truth* or *falsity* of formulas in a world, it is possible to define the *success* or *failure* of events/actions in a world. Rao and Georgeff [107] gives a semantics for events based on their temporal framework but does not address the types of actions involved in particular events. The subtle difference between actions and events lies in the fact that actions are considered to be descriptions of causal processes which upon execution by an agent may turn one state

of affairs into another whereas an event consists of the performance of a particular action by a particular agent. Our aim in this chapter is to investigate the notion of *composite actions* in the framework of BDI thereby answering questions as to:

1. When is it possible for a BDI agent to perform such actions?
2. What are the effects of such an event? (i.e., an event in which a composite action is involved).

This chapter can be viewed as a further extension of the existing BDI theory, whereby, we reason about the mental state of a BDI agent during the execution of an action. The actions are represented in an explicit way in contrast to that of most philosophical accounts [14, 124, 31, 30]. We investigate the close connection between the *result* of an action performed by a BDI agent and its *capability* of achieving that result. We argue that though the agent might have a capability to perform an action it need not be the case that the *opportunity* should always accompany it. This view gets importance when we take into consideration composite actions where one action follows the other $(\pi_1; \pi_2)$, which means an agent performs π_1 followed by π_2. In such cases the *result* of the component parts of the action is needed for the overall success of the action. It also seems reasonable to declare that the agent has the relevant *opportunity* to perform the component actions in such a way that the execution leads to an appropriate state of affairs. By making actions explicit in BDI we try to avoid some of the problems that plague the endogenous systems when dealing with composite actions. We describe a formal relationship between the *result, opportunity, belief, desire* and *intention* modalities. It is important to note the close connection between *intention* and *result*. For instance, if an agent intends to perform a plan, we can infer under certain conditions he intends the result of the plan. Similar is the case with Goals and Results. This work is partially motivated by the KARO architecture of Van Linder [123] whereby we indicate how *result* and *opportunity* can be integrated to the existing BDI framework. Such an addition definitely paves way for a better understanding of the dynamics [48, 104] of BDI Systems.

4.2 Modal Logics of Action and Ability

In this section we give a brief account of some interesting theories pertaining to the modal logics of agency and the related concepts of action and ability. The central idea of these theories is that the notion of agency should be

analysed as a web of concepts pertaining to *control*. *Control* is interpreted as a power to bring about something, i.e., control involves a goal towards which the agent is directed. Such a notion of agency gains importance when viewed in the backdrop of BDI, as the central characteristic of a BDI agent is its *goal-directed behavior*. Though the main theme of these theories and that of BDI remains the same, the difference arises in our attempt to include explicitly the different action constructs which includes composite actions.

4.2.1 Brown's theory of *Intentional* control

Among the different theories we consider, Brown's logic of ability [14] seems to be closely related to our work. Brown's logic of ability is based on classical modal logic, i.e., logics closed under substitution of logical equivalents:

$$\textbf{RE:} \qquad \text{If } \models \varphi \Leftrightarrow \psi \text{ then } \models \Theta\varphi \Leftrightarrow \Theta\psi$$

where Θ stands for the respective *operator*. The main motivation for using classical modal logic is to avoid the schema

$$\textbf{DisAble:} \qquad \mathit{Ability}(\varphi \vee \psi) \Rightarrow (\mathit{Ability}\varphi \vee \mathit{Ability}\psi)$$

which is unwanted in a system where ability is *intention-based* (i.e., ability is interpreted as *intentional control*[1]). Brown gives the following example in this regard:

> even if a man might have the ability to draw a red card or draw a black card from a deck of cards, he might not have the ability to draw a red card or have the ability to draw a black card.

But, as pointed out by Dag Elgesem [31], by using a classical modal logic Brown is not able to avoid **DisAble**. The reason is that if the operator interpreted as ability is defined as *normal*-possibility operator then one can automatically get **DisAble** (**C$\Diamond$:** $\Diamond(\varphi \vee \psi) \Rightarrow (\Diamond\varphi \vee \Diamond\psi)$). Though Brown gives this as a reason for employing a classical modal logic to characterize the logic of ability, at the same time, he claims that one should only use classical necessity operator to model ability. If so then *normal* necessity operator will not have **DisAble** and hence there is no need to adopt a classical system to avoid **DisAble**. There seems to be a contradiction

[1]To say that ability is intention based is to think that intentional repetition is a necessary condition for ability.

in Brown's account here. Another feature of Brown's logic is the use of *minimal models*. A minimal model is a structure $\mathfrak{M} = (\mathsf{W}, f, \nu)$ where W is a set of possible worlds and ν gives a truth value to each atomic sentence at each world and f is a function that associates with each possible world a collection of sets of possible worlds called a *cluster* of worlds. The intuitive idea is that each cluster of worlds represents an action done by the agent, and that the agent could be said to have the ability to do an action φ if there exists a relevant cluster wherein all of the worlds in the cluster φ holds. The importance of this interpretation lies in the fact that ability requires repeatability which in turn points out to an important aspect of control viz. *reliability*. The following truth-definition is given for the notion of ability and the logic is closed under the rules **RE** and **RM**[2]

$$\mathfrak{M}, \omega \models \textbf{Ability}\ \varphi \text{ iff } \exists \mathsf{W}' \subseteq \mathsf{W}, \omega \mathsf{R} \mathsf{W}' \text{ and } \forall \omega' \in \mathsf{W}' : \mathfrak{M}, \omega' \models \varphi$$

It should be noted that in the formalism $\omega \mathsf{R} \mathsf{W}'$ R is defined on worlds. Other notions such as *practical necessity*, *might*, *will* etc. has been explained by Brown in relation with agency. Among these the notion of practical necessity is interesting for analysing agents in planning systems. The importance of formalising the knowledge the agents have on their practical possibilities have been pointed out by Vanlinder [123]. The following definition is given for practical necessity by Brown:

$$\mathfrak{M}, \omega \models \textbf{Will}\varphi \text{ iff } \forall \mathsf{W}' \subseteq \mathsf{W},\ \text{if } \omega \mathsf{R} \mathsf{W}' \text{ then } \forall \omega' \in \mathsf{W}' : \mathfrak{M}, \omega \models \varphi$$

What this means is that a proposition φ is practically necessary for the agent if every world in every cluster has φ true, hence φ is unavoidable for the agent, i.e., in every world where the agent does something φ holds. The notion of practical possibility is obtained by taking the dual, i.e., $\neg\textbf{Will}\neg\varphi$ and is termed **Might** by Brown. The logic of **Will** is closed under **RE**, **RM** and **NEC**. Brown gives a range of axioms that captures the interaction of the different modal operators.

4.2.2 Elgesem's *Non-Intentional* Theory of Ability

Another interesting theory of agency as goal-directedness is the one proposed by Elgesem [31] which is based on the notion of *objective goal-directedness*. The main idea is to analyse goal-directedness as the ability on the part of the agent to maintain a certain goal state in the face of variations in the environment so that its relationship with the environment is

[2] $\models \varphi \Rightarrow \psi$ then $\models \Box\varphi \Rightarrow \Box\psi$.

not destroyed. Elgesem's theory is based on the work of Sommerhoff [118] and adopts a *non-intentional* strategy of agency. Though this conflicts with the BDI theory of agency some of their notions are helpful in explaining our theory of composite actions. By saying that Elgesem's theory is non-intentional we mean that the notion of objective goal directedness involves directedness towards a goal state without the agent necessarily being *aware* of it (i.e., the agent is not aware of this as a state towards which its activity is directed). Three criteria–criteria of **success**, **avoidability** and **non-accidence**– are given central role in this theory. The criteria of Success demands a causal condition for the goal-event to occur, i.e., a particular relationship exists between the agent and the environment which gives a *causally necessary* condition for the goal event to occur. Formally this is given as

$$F(\mathbf{a}_k, \mathbf{e}_k) = 0 \tag{4.1}$$

and is interpreted as follows: For an action which is directed towards a goal $\mathbf{G}$, there exists at-least one point of time t_k during the action and at-least one variable $\mathbf{a}$ associated with the action and at-least one variable $\mathbf{e}$ associated with the environment such that at t_k it is a necessary condition for the occurrence of the goal event $\mathbf{G}$ that $\mathbf{a}_k$ has a specific relation to $\mathbf{e}_k$. This condition is often referred to as the **Focal condition**.

The criteria of avoidability is a negative condition saying that some possible state of the system plus environment could not be a goal-state for the agent if it always obtains in every state of the world. Here, by state, the authors refer to different time points. Finally, the criteria of non-accidence says that the production of the *causally necessary* condition should be the manifestation of a capacity on the part of the agent. This is a counter-factual condition which stresses the fact that

> the agent produces the appropriate action not only under the actual environmental conditions, but also that it would have produced an appropriate modified action under a variety of alternative circumstances each requiring a specific modification of the action [31].

In order to capture semantic conditions of the above theory Elgesem adopts the following formalism:

$$\mathfrak{M} = (\mathsf{W}, f_i, \nu)$$

where W is a set, ν is a valuation function and $f_i = f_1, \ldots, f_n$ are functions from $\mathsf{W} \times \wp(\mathsf{W}) \Rightarrow \wp(\mathsf{W})$ where $\wp(\mathsf{W})$ stands for the *power set* of W. The selection function is defined as

$$f_i(\omega, || \mathbf{G} ||^{\mathfrak{M}})$$

where i is an agent ω is the actual world and $||\ \mathbf{G}\ ||^{\mathfrak{M}}$ is the set of worlds where some goal $\mathbf{G}$ is true. The value of the function $f_i(\omega, ||\ \mathbf{G}\ ||^{\mathfrak{M}})$ is the set of worlds where the agent realises the ability he has in ω to bring about the goal $\mathbf{G}$. For each agent $1, \ldots, n$ one such function is defined. Based on this function the *focal condition* as stated in (4.1) is captured as follows:

> $\omega' \in f_1(\omega, ||\ \mathbf{G}\ ||^{\mathfrak{M}})$if and only if the **Focal condition** obtains in ω',
> *i.e.*, $f'(o_{\omega'}, e_{\omega'}) = 0$ for some function f'

where $o_{\omega'}$ and $e_{\omega'}$ denote the value in ω' of some behavioral and environmental variables. The function $f_i(\omega, X)$ is interpreted as giving the set of worlds where the agent i realises his ability to bring about the goal $\mathbf{G}$. In order to account for the **Success** condition Elgesem introduces the following constraint on the function

$$f_i(\omega, X) \subseteq X \text{ where } X \subseteq \mathsf{W} \tag{4.2}$$

The constraint conveys the idea that in all worlds where the agent exercises his ability to bring about the goal, the goal is realised. It is important to distinguish between what it means to say that the goal is realised in every world where the agent exercises his ability to bring about the goal and that of he in fact brings it about in the actual world. In the case of the later the semantic condition is given as follows:

$$\mathfrak{M}, \omega \models\ \mathit{Does}_i\ \varphi \text{ if and only if } \omega \in f_i(\omega, ||\ \varphi\ ||^{\mathfrak{M}}). \tag{4.3}$$

Based on this notion of *bringing it about* Elgesem goes on to define his concept of ability as given below:

$$\mathfrak{M}, \omega \models\ \mathit{Ability}_i\ \varphi \text{ if and only if } f_i(\omega, ||\ \varphi\ ||^{\mathfrak{M}}) \neq \emptyset. \tag{4.4}$$

The difference between *bringing it about* (4.2) and *ability* (4.3) lies in the fact that in the former the agent is actually being directed towards some goal whereas in the later the agent has the power or ability to realise the goal state. In a similar manner, the criteria of **Success**, as mentioned above, is captured through 4.2 and that of **Non-accidence** through 4.3 and 4.4 as is stated in the following axioms

$$\begin{array}{ll} (A1) & \mathit{Does}_i\varphi \Rightarrow \varphi \\ (A2) & \mathit{Does}_i\varphi \Rightarrow \mathit{Ability}_i\varphi \\ (A3) & \neg\mathit{Ability}_i\top \end{array}$$

A1 comes from 4.2 and A2 from 4.3 and 4.4. The third axiom is related to the **Avoidability** criteria and in-order to account for it Elgesem introduces further constraints on the focal condition as is shown below:

$$f_i(\omega, \mathsf{W}) = \emptyset \tag{4.5}$$

The constraint stresses on the point that the focal condition should not be realisable in every possible world and A3 captures this property by saying that the goal state which is realised in every world is not a state any agent is able to bring about. What this means is that other requirements like the relationship between environmental variables and behaviour variables should be taken into account (Elgesem gives the condition that they should be orthogonal).

Having dealt with all of Somerhoff's criteria, Elgesem goes on to analyse other related theories and explains several other properties that are characteristic of his notion of agency. We do not want to outline the whole theory here but would rather point out the difference in our approach. For further details refer [31]. The fundamental difference is that Elgesem's theory is a non-intentional one, as stated before, and is based on the concept that a theory of agency need not always have an intentional look. But, whereas with BDI one can see that it is heavily based on cognitive notions like belief, desire and intention. Moreover, the fundamental characteristic of a BDI agent is the dynamics involved in the interaction between these three modalities. Hence, if we want to introduce a notion of *ability* it should be in alliance with these mental notions. As far as Brown's theory goes, though he characterises ability as *intentional control*, he fails to relate it with intentional action and in a way is not able to explain the **Non-accident** criteria. It is also the case that neither Brown's nor Elgesem's theory has a notion of composite actions like $\alpha_1; \alpha_2$ (α_1 *followed by* α_2). When we take such actions into consideration the *result* of the first action is important for the successful execution of the second and thereafter for the success of the whole action as such. Moreover, both theories use modal operators as ranging over formulae to formalise ability as opposed to our approach of representing the actions in an explicit manner. For a slightly critical examination of Elgessem's axiomatisation refer [62].

4.2.3 BDI with Capabilities

Recent work by Padgham [97] shows how a notion of capability can be integrated into the BDI logic. Their work is motivated by the use of a *capability* construct in *Jack* [16], a java based BDI agent development environment. Capability is interpreted in terms of plans, and plans are specifications for achieving certain goals. Each plan is associated with a triggering event which often has the form *achieve goal* α. To say that an agent has a *capability to achieve* α means that the agent has at-least one plan that has as its trigger the goal event *Achieve* α, which in turn implies that the agent has at-least one way it knows how to achieve α in some situation. It may be

the case that at any given time the agent is unable to use this plan as each plan has a precondition (context) which describes the situation in which the plan is intended to be used. This precondition should match the state of the world for the agent to perform the plan. Whether the precondition matches the state of the world or not, the important point here is that having such a plan is a prerequisite to being able to achieve α. We do not want to discuss the details regarding the implementation of the *capability* construct in *Jack* but would get into the logical formalism as given by Padgham.

The extended version of BDI logic with *capability* is called the *IC-system*[3] by Padgham. We do not want to outline the basic I-system here as this was done in Chapter 2 but will concentrate on the newly added *capability* construct. As far as the semantics goes capability is given equal status with that of beliefs, desires and intentions. This means that capability is not reduced to any of these notions. CAP(φ) is then defined as being true if it is true in all the capability-accessible worlds . If C is the accessibility relation with respect to capabilities, then the following semantic condition holds:

$$\mathfrak{M}, \omega \models \text{CAP}(\varphi) \text{if and only if } \forall \omega' \in C^{\omega} : \mathfrak{M}, \omega' \models \varphi$$

The **K** and **D** axioms are adopted for capabilities. We do not refer to any temporal notions in this framework. When it comes to *interaction axioms* of the type $\Box_1 \Rightarrow \Box_2$, (originally *compatibility axioms*), the set and structural relationships are used to explain the relation as was shown in chapter 2. The interaction axiom for Belief-Capability and their corresponding structural relationship is given as follows:

$$\text{CAP}(\varphi) \Rightarrow \text{BEL}(\varphi)$$
$$\forall \omega' \in B^{\omega}, \exists \omega'' \in C^{\omega} \text{ such that } \omega'' \sqsubseteq \omega' \ (i.e. \mathsf{B} \subseteq \mathsf{C})$$

The formalism conveys the meaning that if an agent i has a capability to achieve φ, then agent i believes that it is possible for φ to be true. In a similar manner the capability-goal interaction axiom and the corresponding structural relationship is explained as follows:

$$\text{GOAL}(\varphi) \Rightarrow \text{CAP}(\varphi)$$
$$\forall \omega' \in C^{\omega}, \exists \omega'' \in G^{\omega} \text{ such that } \omega'' \sqsubseteq \omega' \ (\mathsf{C} \subseteq \mathsf{G})$$

In addition to these interaction axioms Padgham goes on to define mixed modal axioms and their corresponding semantic conditions. These definitions too are characteristic of the original BDI axioms as was shown in

[3]The original system proposed by Rao and Georgeff [107] is called the I-system

Chapter 2. The mixed modality axiom showing capabilities regarding goals and their corresponding semantic condition is stated as follows,

$$\text{GOAL}(\varphi) \Rightarrow \text{CAP}(\text{GOAL}(\varphi))$$
$$\forall \omega' \in C^{\omega}, \forall \omega'' \in G^{\omega'} \text{ we have } \omega'' \in G^{\omega}$$

with the reading that *If the agent has a goal φ then it has the capability to have the goal φ*. Beliefs about capabilities and capabilities regarding intentions can be given in a similar manner. In section 4.3, while defining our approach, we give a full list of the axioms characterising the I-system of Rao and Georgeff as well as the IC-system of Padgham. Further, based on their IC-system Padgham and Lambrix examines the *commitment* axioms as given in [107] and come up with their notion of a *self-aware* agent. To explain the concept of a self-aware agent we have to check the commitment strategies of an agent as given by Rao and Georgeff. According to them there are three different commitment strategies: *blind, single-minded* and *open-minded*. The blindly committed agent maintains its intentions until they are believed true, the single-minded agent maintains intentions until they are believed true or are believed impossible to achieve, while the open-minded agent maintains intentions until they are believed true or are no longer goals. Compared with this a self-aware agent is one which is able to drop an intention if it believes it no longer has the capability to achieve that intention. An axiom corresponding to this condition together with the IC-system is used to define a self-aware agent. Padgham and Lambrix proves certain properties of their logic and also shows how the notion of capability could be adopted in the abstract BDI interpreter given in [108].

It should be noted that the theories described so far talk about action related constructs and their formalism without directly dealing with actions as such. Though the BDI formalism talks about events they do not mention anything about the actual execution of actions. Moreover none of these theories deals with the notion of composite actions. In the next section we show the subtle difference involved in *intentional* action and *intending* to do an action as outlined in Bratmans' theory and on which we base our explanation of composite actions.

4.2.4 Intentional Action vs Intending an Action

When one takes into account the compositional nature of actions $\alpha_1; \alpha_2$ (α_1 followed by α_2), it seems contradictory to believe that an implicit representation alone could account for the mental state of an agent during the execution of such actions. The problem with the current formalisms is in their failure to differentiate *intentional action* (predefined intention) from

intending to do an action (future intention). The former applies to the action or states that the agent performs or brings about but not with any prior intention to do so whereas the later involves the true intentions or preferences of the agent. Most of the work in BDI represent actions in the former manner. For instance, in the work of Rao [107] axioms like

$$\text{INT}(does(e)) \Rightarrow does(e)$$

are used to capture the volitional commitment of an agent stating that whenever an agent has an intention to a particular primitive action he/she will do that action. The formalism remains true for single actions, but when it comes to composite actions like $(\alpha_1; \alpha_2)$ it fails to do justice as it is taken for granted that the execution of the first action necessarily leads to the second without mentioning anything about the *result* of the first action on the second. Based on the existing BDI architecture the concept of composite actions could be formalised as

$$\text{INT}(does(\pi_1; \pi_2)) \Rightarrow does(\pi_1; \pi_2).$$

This need not be the case as the performance of π_1 could result in a counter-factual state of affairs. It seems crucial to consider the *result* of the first action for the overall success of the composite action. In the same manner formulas like

$$\text{GOAL}(\pi_1; \pi_2) \Rightarrow \text{CAP}(\text{GOAL}(\pi_1; \pi_2))$$

seem to be problematic as the formulation does not tell anything about the ability of the agent if the first action results in a counter-factual state of affairs. It does not mention anything regarding the *opportunity* the agent has in performing the second action.

It is important to make a division between the two notions of of *Intentional* and *Intending* for our framework. The former relates to a predefined intention, where the *Result* of an action is taken for granted, whereas the latter concerns a future intention, where further deliberation is done as to what the result would be before an action is performed. Davidson [124] oversees such a division and extends the concept of *intentionally doing* to that of *intending to*. Though Bratman [12] points out this disparity the current formalisms does not allow for sound representation using the existing modal operators. Hence the need for additional constructs like RES and OPP. In intentional action, there is no temporal interval between what Davidson terms as *all-out evaluation and action*. So there is no room for further practical reasoning in which that all-out evaluation can play a significant role as input. The BDI framework gives primary importance to

practical reasoning and hence to means-end reasoning which is important to avoid further deliberation at the time of action. Therefore it seems appropriate to categorise composite actions under future intentions as they play a crucial role in our practical thinking. More importantly, we form future intentions as part of larger plans whose role is to aid co-ordination of our activities over time. As elements in these plans, future intentions force the formation of yet further intentions and constrain the formation of other intentions and plans.

4.3 Integrating Results

In this section we show how a notion of *Result* could be integrated to the logic of BDI without altering the basic framework. We briefly introduce the axioms corresponding to the I-system as well as the IC-system and show why we need extra constructs to accommodate composite actions. Before going onto the details we need to fix certain definitions.

Definition 29 *Let φ be a formula,* BEL*,* INT *and* GOAL *be the modal operators for the mental constructs, done, does be the operators for event types, and* inevitable *be the modal operator for a path formulae*[4]*; then Table 4.1 gives the corresponding axioms for the I-system and IC-system.*

The I-system		**The IC-system**	
A1	$\text{GOAL}(\varphi) \Rightarrow \text{BEL}(\varphi)$	**C1**	$\text{CAP}(\varphi) \Rightarrow \text{BEL}(\varphi)$
A2	$\text{INT}(\varphi) \Rightarrow \text{GOAL}(\varphi)$	**C2**	$\text{GOAL}(\varphi) \Rightarrow \text{CAP}(\varphi)$
A3	$\text{INT}(\mathit{does}(\pi)) \Rightarrow \mathit{does}(\pi)$	**C4**	$\text{CAP}(\varphi) \Rightarrow \text{BEL}(\text{CAP}(\varphi))$
A4	$\text{INT}(\varphi) \Rightarrow \text{BEL}(\text{INT}(\varphi))$	**C4**	$\text{GOAL}(\varphi) \Rightarrow \text{CAP}(\text{GOAL}(\varphi))$
A5	$\text{GOAL}(\varphi) \Rightarrow \text{BEL}(\text{GOAL}(\varphi))$	**C5**	$\text{INT}(\varphi) \Rightarrow \text{CAP}(\text{INT}(\varphi))$
A6	$\text{INT}(\varphi) \Rightarrow \text{GOAL}(\text{INT}(\varphi))$		
A7	$\mathit{done}(\pi) \Rightarrow \text{BEL}(\mathit{done}(\pi))$		
A8	$\text{INT}(\varphi) \Rightarrow \mathit{inevitable}\Diamond(\neg\text{INT}(\varphi))$		

Table 4.1: Axioms related to the I-system of Rao and IC-system of Padgham

Axiom A3 seems to be problematic because of the fact that the event π need not be necessarily restricted to a single action. If the agent has a

[4]Though we do not use any temporal concepts the notion of path formulae is needed to explain the commitment axioms given in Section 4.6. *state formulas* are formulas that are evaluated at a given time point in a given world whereas *path formulas* are formulas that are evaluated along a given path in a given world.

choice of actions at a particular world, he/she would be incapable of acting *intentionally* until she deliberates and chooses one of them. It is the same case when the particular event is a composite action. The agent needs to deliberate on the *result* of the first action for the successful execution of the second one. It might also be the case that the agent lacks the relevant *opportunity* at that particular world of doing the specific action. The result and opportunity constructs become more relevant in the case of IC-system as given in Table 4.1 as they can convey some intuition regarding the capability of an agent in a counter-factual state of affairs. Axiom A8 has got some temporal constructs like *inevitable*$\Diamond$ (inevitable eventually) and conveys the meaning that *if an agent forms an intention, then some time in the future it will give up that intention.* We wanted to list all the axioms corresponding to the I-system and hence didn't want to avoid A8.

It is clear that from what has been said till now regarding BDI based systems the compositional behaviour of actions has not been dealt within the BDI architecture. With the recent addition of the *Capability* construct we believe that it is worthwhile exploring this concept. Whereas the BDI framework is concerned with finding out what it means exactly to have the ability to perform some action, we try to focus on the compositional behavior of actions. In other terms we are concerned with finding a relation between the capability to perform a composite action and relate it with the capability for the components of that action. Not all actions are treated equally in our approach but instead the result of each action is determined individually and then the conclusion is made whether the agent succeeds in performing that action.

Three types of actions are dealt with $(\pi_1; \pi_2)$ (π_1 followed by π_2), (*while* φ *do* π) (π as long as φ holds) and (*if* φ *then* π_1 *else* π_2) (π_1 if φ holds and π_2 otherwise). The composite action $(\pi_1; \pi_2)$ is discussed in detail. An additional operator RES (*result*) is introduced to show the success/failure of the component actions. The RES operator functions as a *practition* operator which indicates the sequence of actions being performed, i.e., which action is performed next. The existing BDI architecture does not mention anything about the actual execution of actions. Since the transition caused by the execution of the action $(\pi_1; \pi_2)$ equals the *sum* of the transition caused by π_1 and the one caused by π_2 in the state brought about by execution of π_1, the RES operator helps in acting as a filter which checks whether the first action results in a counter-factual state or not. Such a filtering helps in avoiding further deliberation at the time of action as would otherwise be in situations arising from counter-factual states. For example, the success of the printer command (*lpr*), in a Unix environment, depends on the result of the execution of the command in the spooler phase followed by the recogni-

tion of the command by the printer in the communication phase. Here the action needs to be broken down into compartments and the success of each action should be validated for the overall success. In such circumstances the RES operator helps in providing the necessary specification. This goes in alliance with our view of categorising composite actions under future intentions, where the scope of practical reasoning is more.

Definition 30 *Let π_1, π_2 be actions, then the axioms for the operator* RES *are:*

R1 $\text{CAP}(does(\pi_1;\pi_2)) \Rightarrow \bigwedge_{i=1,2} \text{BEL}(does(\pi_i)) \wedge \text{BEL}(\text{RES}(does(\pi_1)) \neq \bot)$

R2 $\text{GOAL}(does(\pi_1;\pi_2)) \Rightarrow \bigwedge_{i=1,2} \text{CAP}(does(\pi_i)) \wedge \text{RES}(does(\pi_1) \neq \bot)$

R3 $\text{CAP}(does\pi_1;\pi_2)) \Rightarrow \bigwedge_{i=1,2} \text{BEL}(\text{CAP}(does(\phi_i))) \wedge \text{RES}(does(\pi_1) \neq \bot)$

R4 $\text{GOAL}(does(\pi_1;\pi_2)) \Rightarrow \bigwedge_{i=1,2} \text{CAP}(\text{GOAL}(does(\pi_i))) \wedge \text{RES}(does(\pi_1) \neq \bot)$

R5 $\text{INT}(does(\pi_1;\pi_2)) \Rightarrow \bigwedge_{i=1,2} \text{CAP}(\text{INT}(does(\pi_i))) \wedge \text{RES}(does(\pi_1) \neq \bot)$

The first axiom states that an agent has the capability of performing a composite action $\pi_1;\pi_2$ then at some point of time the agent believes in doing π_1 and π_2 and believes that the performance of π_1 does not end in counter-factual state of affairs (i.e, it does not end in falsity). Similarly the third axiom states that an agent has the capability of performing a composite action $\pi_1;\pi_2$, then at some point of time, the agent believes that it has the capability of doing π_1 and believes in the capability of doing π_2 and the result of π_1 does not end up in a counter-factual state of affairs.

The semantic conditions for RES is the same as given for CAP in section 4.2.3. For instance it can be shown that the semantic condition for R2 is

$$\forall w' \in C_t^w(i), \exists w'' \in G_t^w(i), \exists w''' \in R_t^w(i) \text{ such that } w'' \sqsubseteq w' \text{ and } w''' \sqsubseteq w'$$

where $R_t^w(i)$ is the set of result-accessible worlds of agent i in world w at time t. The temporal variable t is static as we do not make any explicit representation of time. This constraint means that for every capability-accessible world w' at time-point t, there is a goal-accessible world w'' at

that time-point which is a sub-world of w' and a result-accessible world w''' which is a sub-world of w'. The converse doesn't hold as there can be Goal-accessible worlds that do not have corresponding capability as well as result-accessible worlds that do not have corresponding capability but only has the opportunity. We shall deal with the opportunity construct in the next section.

The action constructors dealing with *while* φ *do* π (which means that π as long as φ holds) and *if* φ *then* π_1 *else* π_2 (π_1 if φ holds and π_2 otherwise) is crucial from computational point of view. For an agent to be able to perform an action *while* φ *do* π, it has to have the ability to perform some finite actions constituting the body of the while-loop as well as the opportunity to perform all the steps. Agents should not be able to perform an action that goes indefinitely. These specifications are formally represented by the following two axioms.

R6 $\text{CAP}(\textit{while}\,\varphi\,\textit{do}\,\pi) \Rightarrow [\neg\varphi \vee (\varphi \wedge \text{BEL}(\text{CAP}(\textit{does}(\pi)))) \wedge \text{RES}(\textit{done}(\pi)) \neq \bot]$

R7 $\text{CAP}(\textit{if}\,\varphi\,\textit{then}\,\pi_1\,\textit{else}\,\pi_2) \Rightarrow$
$[\varphi \wedge \text{BEL}(\text{CAP}(\textit{does}(\pi_1))) \wedge \text{RES}(\textit{done}(\pi_1)) \neq \bot] \vee$
$[\neg\varphi \wedge \text{BEL}(\text{CAP}(\textit{does}(\pi_2))) \wedge \text{RES}(\textit{done}(\pi_2)) \neq \bot].$

The first proposition states that an agent is capable of performing an action *while* ϕ *do* π, as long as ϕ holds and the agent believes that it has the capability of doing π and result of π does not end in falsity. Similarly R7 can be read as, an agent has the capability of performing an action *if* ϕ *then* π_1 *else* π_2, if ϕ holds and the agent believes that it has the capability of doing π_1 and the result of π_1 is true, or it is the case that, ϕ does not hold and the agent believes that it has the capability of doing π_2 and result of π_2 does not end in a counter-factual state of affairs.

4.4 Integrating Opportunity

Though in many cases it seems reasonable to assume that Capability implies Opportunity, when it comes to practical reasoning Opportunity seem to play a significant role. Van Linder [123] explains opportunity in terms of the correctness of action. An action is correct for some agent to bring about some proposition iff the agent has the opportunity to perform the action in such a way that its performance results in the proposition being true. Integrating opportunity lays further constraint on the part of the agent to think about an action before getting committed. Consider the

example of a lion in a cage, which is perfectly well capable of eating a zebra, but ideally never has the opportunity to do so.[5] Using the BDI formalism we would have to conclude that the lion is capable of performing the sequential composition *eat zebra* ; *fly to the moon* which hardly seems to be intuitive. In such situations it is very important to know the combination of Capability and Opportunity so that no unwarranted conclusions can be drawn. We introduce an operator OPP whose intuitive meaning is agent x has the opportunity. The axioms for the OPP operator together with the Capability construct can be given as follows

Definition 31 *Let π_1, π_2 be actions, then we have*

O1 $\mathrm{CAP}(does(\pi_1;\pi_2)) \Rightarrow \bigwedge_{i=1,2} \mathrm{BEL}(\mathrm{OPP}(does(\pi_i)))$

O2 $\mathrm{GOAL}(does(\pi_1;\pi_2)) \Rightarrow \bigwedge_{i=1,2} \mathrm{CAP}(does(\pi_i)) \wedge \mathrm{OPP}(does(\pi_i))$

O3 $\mathrm{CAP}(does(\pi_1;\pi_2)) \Rightarrow \bigwedge_{i=1,2} \mathrm{BEL}(\mathrm{CAP}(does(\pi_i))) \wedge \mathrm{OPP}(does(\pi_i))$

O4 $\mathrm{GOAL}(does(\pi_1;\pi_2)) \Rightarrow \bigwedge_{i=1,2} \mathrm{CAP}(\mathrm{GOAL}(does(\pi_i))) \wedge \mathrm{OPP}(does(\pi_i))$

O5 $\mathrm{INT}(does(\pi_1;\pi_2) \Rightarrow \bigwedge_{i=1,2} \mathrm{CAP}(\mathrm{INT}(does(\pi_i))) \wedge \mathrm{OPP}(does(\pi_i))$

O6 $\mathrm{CAP}(\mathit{while}\,\varphi\,\mathit{do}\,\pi\,) \Rightarrow [\neg\varphi \vee (\varphi \wedge \mathrm{BEL}(\mathrm{CAP}(does(\pi)))) \wedge \mathrm{OPP}(does(\pi))]$

O7 $\mathrm{CAP}(\mathit{if}\,\varphi\,\mathit{then}\,\pi_1\,\mathit{else}\,\pi_2) \Rightarrow [\varphi \wedge \mathrm{BEL}(\mathrm{CAP}(does(\pi_1))) \wedge \mathrm{OPP}(does(\pi_1))] \vee$

$[\neg\varphi \wedge \mathrm{BEL}(\mathrm{CAP}(does(\pi_2))) \wedge \mathrm{OPP}(does(\pi_2))]$

The third axiom states that an agent has the capability of performing $\pi_1;\pi_2$ then the agent believes that he has the capability of ϕ_1, if he has the opportunity of doing π_1, and, he has the capability of π_2, if he has the opportunity of π_2. Similarly O7 can be interpreted as an agent has the capability of doing the action (if φ then π_1 else π_2) then either φ holds and the agent believes that he/she has the capability of doing π_1 provided the opportunity exists or $\neg\varphi$ holds and the agent has the capability of doing π_2 provided the opportunity exists. The other axioms can be interpreted in a similar manner.

[5] The example is taken from [123].

4.5 Opportunity + Results

In [123] a division is made between *optimistic* and *pessimistic agents* and the interpretation of the OPP formulae is done accordingly. They make use of two dynamic operators $\langle do_i(\alpha)\rangle\varphi$ and $[do_i(\alpha)]\varphi$. The first one denotes that an agent i has to have the opportunity to perform the action α in such a way that φ will result from the performance (*Pessimistic Approach*): A pessimistic agent needs certainty. The second one is the dual of the first and states that if the opportunity to do α is present then φ would be among the results of $do_i(\alpha)$ (*Optimistic Approach*). The formula $[do_i(\alpha)]\varphi$ is noncommittal about the opportunity of the agent i to perform the action α. We do not go for such a division and interpret the OPP formulae in a realistic manner linked with the RES operator. Such a formalism helps in avoiding unwarranted results as were seen in the earlier examples. In what follows we present the axioms capturing this intuition.

OR1 $\mathrm{CAP}(does(\pi_1;\pi_2)) \Rightarrow$
$$\bigwedge_{i=1,2} [\mathrm{BEL}(does(\pi_i)) \wedge \mathrm{OPP}(does(\pi_i))] \wedge \mathrm{RES}(done(\pi_1)) \neq \bot$$

OR2 $\mathrm{GOAL}(does(\pi_1;\pi_2)) \Rightarrow$
$$\bigwedge_{i=1,2} [\mathrm{CAP}(does(\pi_i)) \wedge \mathrm{OPP}(does(\pi_i))] \wedge \mathrm{RES}(done(\pi_1)) \neq \bot$$

OR3 $\mathrm{CAP}(does(\pi_1;\phi_2)) \Rightarrow$
$$\bigwedge_{i=1,2} [\mathrm{BEL}(\mathrm{CAP}(does(\pi_i))) \wedge \mathrm{OPP}(does(\pi_i))] \wedge \mathrm{RES}(done(\pi_1)) \neq \bot$$

OR4 $\mathrm{GOAL}(does(\pi_1;\pi_2)) \Rightarrow$
$$\bigwedge_{i=1,2} [\mathrm{CAP}(\mathrm{GOAL}(does(\pi_i))) \wedge \mathrm{OPP}(does(\pi_i))] \wedge \mathrm{RES}(done(\pi_1)) \neq \bot$$

OR5 $\mathrm{INT}(does(\pi_1;\pi_2)) \Rightarrow$
$$\bigwedge_{i=1,2} [\mathrm{CAP}(\mathrm{INT}(does(\pi_i))) \wedge \mathrm{OPP}(does(\pi_i))] \wedge \mathrm{RES}(done(\pi_1)) \neq \bot$$

OR6 $\mathrm{CAP}(while\ \varphi\ do\ \pi) \Rightarrow$

$$\left[\begin{array}{c} (\varphi \wedge \mathrm{BEL}(\mathrm{CAP}(does(\pi)) \wedge \mathrm{OPP}(does(\pi)) \\ \wedge\ \mathrm{RES}(done(\pi)) \neq \bot) \vee (\neg\varphi) \end{array}\right]$$

OR7 $\mathrm{CAP}(\mathit{if}\ \varphi\ \mathit{then}\ \pi_1 \mathit{else}\ \pi_2) \Rightarrow$

$$\left[\begin{array}{c} \varphi \wedge \mathrm{BEL}(\mathrm{CAP}(\mathit{does}(\pi_1)) \wedge \mathrm{OPP}(\mathit{does}(\pi_1)) \wedge \\ \mathrm{RES}(\mathit{done}(\pi_1)) \neq \bot) \vee \\ (\neg\varphi \wedge \mathrm{BEL}(\mathrm{CAP}(\mathit{does}(\pi_2)) \wedge \mathrm{OPP}(\mathit{does}(\pi_2)) \wedge \\ \mathrm{RES}(\mathit{done}(\pi_1)) \neq \bot) \end{array}\right]$$

Axioms OR1–OR7 are a formalisation of the results and opportunities together with the capability operator for composite actions. OR3 states that agents have the capability of doing a composite action $(\pi_1; \pi_2)$ to achieve φ then the agent believes that it has the capability, provided the right opportunity, in each of the atomic states and the resulting condition is in alliance with its beliefs, i.e., it does not result in counter-factual situations. The actual execution of actions is made explicit through such a formalisation. Similarly OR6 states that if an agent has the Capability and Opportunity to perform a while-loop then it keeps this opportunity under execution of the body of the loop as long as the condition holds, i.e., as long as the result is true.

4.6 Commitment Axioms Revisited

In Section 4.2.3 we outlined the commitment strategies of an agent, categorising an agent as *blindly committed agent*, *single minded agent*, *open-minded agent* and the recent *self-aware agent*. In this section we show how to accommodate the RES and OPP constructs in the commitment axioms. Though we didn't make use of temporal operators till now, the commitment axioms as given in [107] are heavily based on them. To be on par with the original formalism we adopt the temporal constructs as defined below: There are two path operators *inevitable* and *optional*. A path formula φ is said to be optional if at a particular point in a time tree[6], φ is true of at-least one path emanating from that point; and it is said to be inevitable if φ is true on all paths emanating from that point. The standard temporal operators $\Diamond$ (eventually), $\Box$ (always), $\bigcirc$ (next) and $\bigcup$ (until), operate over sate and path formulae. Using these constructs together with that of BDI we can formalise *blindly committed agent*, *single minded agent* and *open-minded agent* as follows:

CA1 $\mathrm{INT}(\mathit{inevitable}\Diamond\varphi) \Rightarrow$
$\mathit{inevitable}(\mathrm{INT}(\mathit{inevitable}\Diamond\varphi) \bigcup \mathrm{BEL}(\mathrm{RES}(\varphi)))$

[6] As noted earlier Rao and georgeff adopt CTL^* [33] as the underlying temporal logic in BDI.

CA2 $\text{INT}(\textit{inevitable}\Diamond\varphi) \Rightarrow$
$\textit{inevitable}(\text{INT}(\textit{inevitable}\Diamond\varphi) \bigcup \text{BEL}(\text{CAP}(\varphi)) \vee \neg\text{BEL}(\text{OPP}(\textit{optional}\Diamond\varphi)))$

CA3 $\text{INT}(\textit{inevitable}\Diamond\varphi) \Rightarrow$
$\textit{inevitable}(\text{INT}(\textit{inevitable}\Diamond\varphi) \bigcup \text{BEL}(\text{GOAL}(\varphi)) \vee \neg\text{CAP}(\textit{optional}\Diamond\varphi))$

The axiom for *self-aware agent* [97] given below can be added to the above set of commitment axioms directly.

SA1 $\text{INT}(\textit{inevitable}\Diamond\varphi) \Rightarrow$
$\textit{inevitable}(\text{INT}(\textit{inevitable}\Diamond\varphi) \bigcup (\text{BEL}(\varphi) \vee \neg\text{CAP}(\textit{optional}\Diamond\varphi))$

It seems that the formalisation depicted above is much more intuitive than the one given by Rao and Georgeff [107]. For instance the axiom of blind commitment states that, if an agent intends that inevitably φ be eventually true, then the agent will inevitably maintain its intentions until she believes in the *result* of φ. The addition of *result* is important in the sense that the blindly committed agent maintains the intentions until the agent *actually* believes that she has achieved them, i.e., until the agent has a justified true belief. This condition is needed for an agent blindly-committed to her means to inevitably eventually *believe* that she has achieved her means or ends. It also seems to be in alliance with the philosophical theories concerning the nature of belief[7]. Similarly a single-minded agent maintains her intentions as long as she believes that she has got the capability for it. Since we do not say anything about an agent *optionally* achieving particular means or ends, even if the opportunity is present, the agent does not believe that optionally φ be eventually true which is captured by the $\neg\text{BEL}(\text{OPP}(\textit{optional}\Diamond\varphi))$. Finally, an open-minded agent maintains her intentions as long these intentions are still her goals or as long as she lacks the ability of optionally achieving them.

4.7 Summary and Discussion

The representation and reasoning about composite actions in a BDI environment forms the primary contribution of this chapter. Our work is motivated by the fact that many BDI systems provide no clue as to the actual execution of actions, and are only able to perform actions in an endogenous manner. When dealing with composite actions the actual execution of actions need to be represented and reasoned about for the overall success of

[7]Plato seems to be considering some such definition in *Theaetetus* and perhaps accepting one at *Meno*.

the action. The addition of the two operators RES and OPP strengthens the semantics and functions as a filter in avoiding counter-factual situations. Though some mention has been done in [20] about the composite action construct $(\pi_1; \pi_2)$, it has been restricted to the Intention domain and nothing has been mentioned regarding the result of the actions. The only other comparable work is given by [123].

An explicit representation of temporal constructs can be seen as a further extension to this work. We have used the temporal operators only for the commitment axioms. When it comes to composite actions it is important to mention explicitly the time of each action and the temporal duration of the commitment an agent has towards each action. Our earlier work on neighbourhood logic [99] would be appropriate in specifying this criteria. The interpretation of the $\bigcirc$ (next) operator in the original logic needs to be verified. For example when it comes to composite actions like $(\pi_1; \pi_2)$ the temporal operator $\bigcirc$ can be interpreted either as $\Diamond(\phi \Rightarrow \bigcirc\psi)$ or $(\phi \Rightarrow \bigcirc\psi)$. The temporal notion as to whether the action is performed now or eventually needs to be clarified. It would also be worthwhile to investigate $does(\pi_1; \pi_2)$ in terms of $(done(\pi_1); does(\pi_2)$, i.e., to find whether $does(\pi_1; \pi_2)$ is concurrent or sequential.

Chapter 5

Intention Reasoning as Defeasible Reasoning

Upto now, common sense has been under-valued out of misunderstanding, especially by those who are not well endowed with it.

Alfred Adler

Chapter 3 showed how BDI-like logics could be combined through *fibring* and Chapter 4 introduced the concept of *composite action* in a BDI framework. In this chapter we slightly deviate from the original work and argue that the *intention* component of a BDI agent does not always need to be monotonic. Most of the theories formalising intention interpret it as a unary modal operator in Kripkean semantics, which gives it a monotonic look. We argue that policy-based intentions [12] exhibit non-monotonic behavior which could be captured through a non-monotonic system like defeasible logic. To this end we outline a defeasible logic of intention. The problem of *logical omniscience* which usually accompanies normal modal logics is avoided to a great extend in the proposed framework. The proof theory given shows how our approach helps in the maintenance of intention-consistency in agent systems like BDI.

5.1 Background and Motivations

Formalising cognitive states like intention has received much attention within the AI community [25, 107, 117, 133]. All these theories are based on Nor-

mal Modal Logics (NMLs), where intention is formalised into a modal operator on the framework of kripkean possible world semantics. Due to this restriction, these theories suffer from the *logical-omniscience* problem [67, 123]. One of the solutions suggested to overcome this problem is to adopt a non-Kripkean semantics as proposed in [23]. In that work intention is interpreted in terms of its *content* and the intention consequence relation is explained based on the content of two intentions. There is also a *representationalist* theory of intention [74] that employs the minimal model semantics [22] to interpret the intention operator. Work has also been done relating intention to preferences [120] as well as commitments [24, 105]. However none of these theories have explicitly addressed the need for a non-monotonic theory of intention and we argue that to capture the properties involved in *policy-based* intention we need such a non-monotonic setup.

Our claim is based on Bratman's [12] classification of intention as *deliberative, non-deliberative, policy-based* and we show that the notion of *policy-based* intention is a non-monotonic one (i.e., it has a defeasible nature). Though, many of the theories mentioned above are based on Bratman's work, they fail to recognize the non-monotonic component involved in intention. In this Chapter we adopt a particular non-monotonic system (*defeasible logic*) to study the properties involved in *policy-based* intentions and show how one can relate it with an intentional system like BDI [107]. The reason for defeasible logic is due to its computational efficiency [87] and easy implementation [88]. We are unaware of any existing work relating reasoning about intention with non-monotonic reasoning to the best of our knowledge. We believe that our approach helps in bridging the gap between non-monotonic reasoning and reasoning about intention.

The proposed method provides a partial solution to the problem of *logical-omniscience* which usually accompanies intention-formalisms based on normal modal logics. The use of non-monotonic logics in intention reasoning allows the agent to reason with partial knowledge without having a complete knowledge of the environment. This also helps the agent in avoiding a complete knowledge of the consequences. Moreover, we outline a proof-theory whereby one can reason about ways of maintaining intention-consistency in agent systems like BDI. This is important as it helps in understanding the *dynamics* of BDI systems as outlined in [52]. The new approach facilitates the designer of an agent system like BDI in describing rules for constructing Intentions from Goals and Goals from Knowledge. This is in agreement with the *commitment* axioms of Rao and Georgeff [107] and also provides an explanation on the practical nature of intentional systems like BDI.

5.2 Logics of Intention: An overview

In this section we outline two different logical approaches for modeling intentions as given in the literature. The first one is syntax based whereas the second one is a semantic based approach. It is noted that none of these theories explicitly address the need for a non-monotonic theory of intention.

5.2.1 Konolige and Pollack's Theory of Intention

Kurt Konolige and Martha Pollack (K&P) [74] advocates a non-normal modal theory of intention. They argue that normal modal logics (NMLs) are not suitable for modeling intentions as they suffer from the *side-effect problem* [12]. The side-effect problem as stated in Chapter 2 has three variants and is given as follows

$$\models \mathrm{BEL}(\varphi \Rightarrow \psi) \Rightarrow \models \mathrm{INT}(\varphi) \Rightarrow \mathrm{INT}(\psi) \tag{5.1}$$

$$\models \varphi \Rightarrow \psi \Rightarrow \models \mathrm{INT}(\varphi) \Rightarrow \mathrm{INT}(\psi) \tag{5.2}$$

$$\models \varphi \Leftrightarrow \psi \Rightarrow \models \mathrm{INT}(\varphi) \Leftrightarrow \mathrm{INT}(\psi) \tag{5.3}$$

(5.1) is the case of side-effect under belief implications and states that agents intentions are closed under its belief implications. The side-effect under logical consequences is captured by (5.2) and conveys the meaning that an agent's intentions are closed under tautological implications. Similarly (5.3) is the side-effect under logical equivalence. Another negative remark on NMLs is that they do not provide a means of relating intentions to each other which is necessary to describe the means-end connection between intentions. In order to overcome these problems K&P adopts a representationalist theory of intention in the sense that its semantic objects provide a more direct representation of cognitive state of the intending agent. The representationalist part of the model comes in representing the mental state of the agent using *scenarios*. The concept of a *scenario* for a proposition φ, in a language $\mathcal{L}$ is the set of worlds in W that make φ true. Formally

$$\mathfrak{M}_{\varphi} = \{\omega \in \mathsf{W} \mid \omega, \mathsf{W} \models \varphi\}$$

where a scenario for φ ($\mathfrak{M}_{\varphi}$) identifies φ with the subset of W that make φ true. The main idea behind scenarios is to make a division between wanted and unwanted worlds which is based on the notion that intentions divide the possible futures into those that the agent wants or prefers and those he does not. Hence, according to K&P model the interpretation rule for intention must take into account the complement of the intended

worlds and that makes intention a non-normal modal operator. In the model, courses of actions are represented by possible worlds as the concept of intention is intimately connected with choosing among courses of future actions. Each possible world is a complete history, specifying states of the world at all instances of time. As usual there is a distinguished world (the *actual world*) in all worlds that is the evaluation point for statements. Hence the underlying language $\mathcal{L}$ consists of W (set of possible worlds) and for each world $\omega \in \mathsf{W}$ an evaluation function that determines the value of sentences in the language $\mathcal{L}$. For any sentence φ of $\mathcal{L}$, $\omega(\varphi)$ is the truth value of φ. $\mathcal{L}$ is extended to $\mathcal{L}_\Box$ which includes the modal operators $\Box$ and $\Diamond$ in order to talk about contingent and necessary facts. $\Diamond\varphi$ says that there is a world in W for which φ is true and is used to specify the background of physically possible worlds under which reasoning about intentions take place.

$$\omega, \mathsf{W} \models \Diamond\varphi \text{ iff } \exists\omega' \in \mathsf{W} \text{ such that } \omega', \mathsf{W} \models \varphi$$

This is important in describing the structure of a given domain. $\Box$ is defined accordingly. An assumption is made in the K&P model with regard to *primary intentions*, saying that they do not depend on any other intention the agent currently has.

The second component in the K&P model in addition to the possible worlds is that of *cognitive structures*. A cognitive structure consists of the background set of worlds, and the beliefs and intentions of an agent and is defined as follows:

Definition 32 *Cognitive Structure: A cognitive structure is a tuple (*W*,* Σ*, I) consisting of a set of possible worlds* W*, a subset* Σ *of* W *(the beliefs of the agent) and a set I of scenarios over* W *(the intentions of the agent).*

The beliefs of an agent are taken to be the sentences true in all possible worlds of Σ (though this enforces the condition of *Logical omniscience* on the agent's belief's). If we consider Σ as a set of sentences in $\mathcal{L}_\Box$ and $\mathfrak{M}_\Sigma$ the corresponding possible worlds set then the following condition holds in $\mathcal{L}_I$ (the language of intention corresponding to K&P model):

$$(\mathsf{W}, \Sigma, I) \models \text{BEL}(\varphi) \text{ iff } \forall\omega' \in \mathfrak{M}_\Sigma, \omega', \mathsf{W} \models \varphi, (i.e., \mathfrak{M}_\Sigma \subseteq \mathfrak{M}_\varphi)$$

Therefore, the beliefs of an agent are always possible, that is, they are a subset of the possible worlds ($\Sigma \subseteq \mathsf{W}$). This also means that an agent cannot be wrong about necessary truths. The key concept of this model is that intentions are represented with respect to a background of beliefs about possible courses of events (represented by $\Diamond$), as well as beliefs about

contingent facts (represented by BEL). The following are theorems in $\mathcal{L}_I$

$$\text{BEL}(\varphi) \Rightarrow \Diamond\varphi$$
$$\text{BEL}(\Box\varphi) \Leftrightarrow \Box\varphi$$

In order to define intention structure K&P introduces the constraint that each primary intention will include a scenario, i.e., an intention structure is a set of scenarios $\mathfrak{M}_\varphi$. Hence if I is a set of sentences in $\mathcal{L}_\Box$ where each sentence φ stands for its scenario $\mathfrak{M}_\varphi$ then the following holds

$(\mathsf{W}, \Sigma, I) \models \mathbf{I}(\varphi)$ iff $\exists\psi \in I$ such that $\mathfrak{M}_\psi$ is a scenario for $\varphi, i.e. \mathfrak{M}_\psi = \mathfrak{M}_\varphi$.

The reason for identifying intentions with scenarios is to explicitly encode in the semantics the distinction between preferred and rejected possible worlds. From this definition it is easy to show that $\mathbf{I}(\varphi)$ will hold just in case φ is equivalent to some proposition $\psi \in$ INT, given the background structure W.

Proposition 2 *For any structure (*W*,* Σ*, I),*

$$(\mathsf{W}, \Sigma, I) \models \boldsymbol{I}(\varphi) \textit{ iff } \exists\psi \in I, \mathsf{W} \models \Box(\varphi \equiv \psi).$$

K&P shows how their model is equivalent to the minimal modal semantics of Chellas [22]. In order to show that the **I** operator is not subject to closure under logical consequence (5.2) or under the agent's beliefs (5.1) K&P gives the following example. Let $(\mathsf{W}, \Sigma, \{\varphi\})$, be the cognitive structure representing that an agent has the single intention to perform φ. It is assumed that φ logically implies ψ ($\varphi \Rightarrow \psi$) and not the converse. Hence the following condition holds

$$\mathsf{W} \models \Box(\varphi \Rightarrow \psi) \wedge \Diamond(\psi \wedge \neg\varphi).$$

from which it can be shown that $\mathfrak{M}_\varphi \neq \mathfrak{M}_\psi$ because there is a world in which ψ is true but φ is not. Further from the semantics of **I** it can be shown that

$$(\mathsf{W}, \Sigma, \{\varphi\}) \models \mathbf{I}(\varphi) \wedge \neg\mathbf{I}(\psi))$$

The basic idea here is that in order to distinguish the intention of φ from its necessary consequence ψ, there must be at least one possible world in which ψ is true but φ is not. In a similar manner K&P go on to discuss other satisfiability conditions for intention of which the one for conjunction is given below.

$$(\mathsf{W}, \Sigma, \{\varphi \wedge \psi\}) \models \mathbf{I}(\varphi \wedge \psi) \wedge \neg\mathbf{I}(\varphi) \wedge \neg\mathbf{I}(\psi)$$
$$(\mathsf{W}, \Sigma, \{\varphi, \psi\}) \models \mathbf{I}(\varphi) \wedge \mathbf{I}(\psi) \wedge \neg\mathbf{I}(\varphi \wedge \psi).$$

In order to show the relationship between intentions and beliefs K&P introduce notions like *achievability*, *non-triviality* etc. which are captured by imposing the following conditions on cognitive structures:

$$\exists\omega \in \Sigma, \forall\varphi \in I, \omega \in \mathfrak{M}_{\varphi}$$
$$\forall\varphi \in I, \exists\omega \in \Sigma, \omega \notin \mathfrak{M}_{\varphi}$$

From what has been said till now it can be shown that K&P's theory provides solution to (5.1) and (5.2) by making the constraint that if any intention is derivable from any other in the theory then their content must be logical equivalence. But there is no mention about (5.3). Though (5.3) is usually considered harmless, as far as bounded rationality and resource-boundedness is concerned it is inappropriate. For example from

$$\mathbf{I}(\varphi) \Leftrightarrow \mathbf{I}((\varphi \wedge \psi) \vee (\varphi \wedge \neg\psi))$$

the side-effect ψ is introduced conveying the meaning that logical equivalencies are not *cognitive equivalencies*.

5.2.2 Georgeff and Rao's Theory of Intention

Georgeff and Rao (G&R) [52] advocate a *dynamic* theory of intention wherein the main emphasis is in formalising revision of intentions, beliefs and goals. Their main idea is to solve two related problems of logical omniscience, often known as problem of *unrestricted combining* and problem of *unrestricted weakening*, in relation to intention maintenance. The problems are stated as follows:

$$\models (\mathrm{INT}\varphi \wedge \mathrm{INT}\psi) \Rightarrow \mathrm{INT}(\varphi \wedge \psi) \tag{5.4}$$

$$\models \mathrm{INT}\varphi \Rightarrow \mathrm{INT}(\varphi \vee \psi) \tag{5.5}$$

Although Konolige and Pollack's theory as stated in the previous section accounts for (5.4) they do it in relation to static rather than dynamic properties. As noted in Chapter 2 the semantic model of G&R is based on a branching time logic (CTL^*) and consists of sets of possible worlds where each possible world is a branching tree structure with a single past. A particular index within a possible world is called a time point or situation. The branches within a tree structure represent different courses of action or execution paths. Beliefs, desires and intentions are modeled as sets of possible worlds such that corresponding to each belief-accessible world there is a goal-accessible world and an intention-accessible world. These represent the goals and intentions of the agent with respect to that particular belief

world. Each path within the goal accessible world represents an execution path that the agent wants to achieve and each path within the intention accessible world represents an execution path the agent has decided upon. The only constraint being placed is that of intention paths being a subset of both belief and goal paths in terms of the structural relationship as explained in Chapter 2. This is shown in Figure 5.1. This reflects the intu-

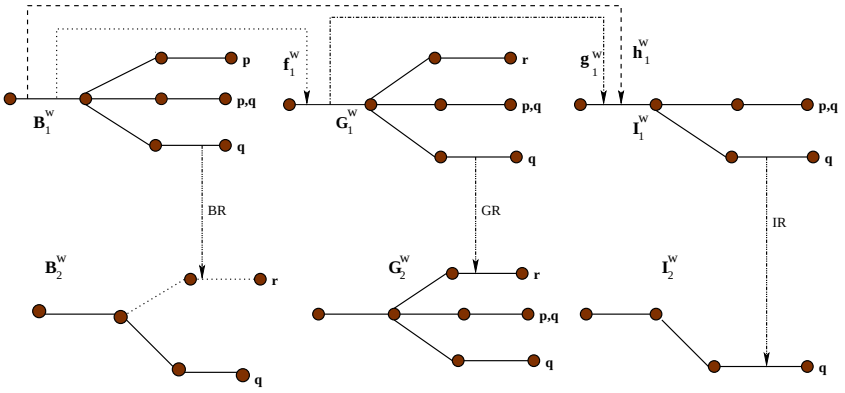


Figure 5.1: An example of belief, desire and intention revision

ition that an agent will only intend a course of action that is both believed possible and has it as a goal. For example in Figure 5.1 as one moves from I_1^ω(intention world at time point 1) to I_2^ω (intention worlds at time point 2) the agents intentions are maintained as long as the agent's beliefs and goals are not significantly changed. Hence the intuition for G&R's formalism is to retain any existing intention path provided it was still both believed possible and the agent has it as its goal. Any intention path that was no longer believed possible or doesn't qualify as goal is pruned off the intention structure. New belief paths, i.e., new opportunities (shown as dotted path with r true in the future, B_2^ω) are not considered. Therefore, all intention paths consistent with satisfying the static structural constraints at the next time point are maintained. The logic used to capture the above intuition is the same as we used in Chapter 2 (section 2.2.1 and 2.2.2) with the addition of three functions f_1^ω, g_1^ω and h_1^ω as shown in Figure 5.1. f_1^ω is a mapping from belief to goal worlds (f_1^ω: $B_1^\omega \Rightarrow G_1^\omega$), g_1^ω is a mapping from goal worlds to intention worlds (g_1^ω: $G_1^\omega \Rightarrow I_1^\omega$) and h_1^ω is a mapping from Belief worlds to intention worlds (h_1^ω: $B_1^\omega \Rightarrow I_1^\omega$). A *deterministic world* assumption is made which requires for a given model and all world time point pairs, that $\forall b, b' \in B_t^\omega$, if $\nu(b,t) = \nu(b',t)$ then $b = b'$ where ν is the truth assignment function. This means that the mappings f, g and h are uniquely determined by the truth function assignment ν, given the assumption of a deterministic

world. The constraints and axioms given by G&R regarding the structural relationship among the modalities is shown in Table 5.1.

Constraints	Axiom
$\forall\omega, \forall t, f_t^\omega, g_t^\omega, h_t^\omega$ are total 1-1 mappings	$\text{INT}(\varphi) \Rightarrow \text{BEL}(\varphi)$
$\forall b \in B_t^\omega \ paths(h_t^\omega(b)) \ \subseteq paths(b)$	$\text{BEL}(\psi) \Rightarrow \text{INT}(\psi)$
$\forall g \in G_t^\omega \ paths(g_t^\omega(b)) \ \subseteq paths(g)$	$\text{INT}(\varphi) \Rightarrow \text{GOAL}(\varphi)$
$\forall b \in B_t^\omega \ paths(g) \ \cap paths(f_t^\omega(b)) \neq \emptyset$	$\text{GOAL}(\psi) \Rightarrow \text{INT}(\psi)$
	$\text{BEL}(\varphi) \Rightarrow \neg\text{GOAL}(\neg\varphi)$

Table 5.1: Structural relationship among modalities in G&R formalism

In order to capture intention maintenance G&R introduce *only* forms of the modalities for beliefs (OBEL), goals (OGOAL) and intentions (OINT) as well as belief revision (BR_t^ω), goal revision (GR_t^ω) and intention revision (IR_t^ω) functions for each world ω and time t. Intuitively, if an agent *only intends* a formula φ then φ is true in all the intention-accessible worlds and the set of intention accessible worlds includes all worlds where φ is true. Formally,

$$\mathfrak{M}, \omega_t \models \text{OINT}(\varphi) \text{ iff } \forall\omega' \in \mathsf{W}, \omega' \in I_t^\omega \text{ iff } \mathfrak{M}, \omega_t' \models \varphi$$

G&R prove that OINT is not closed under unrestricted combination and gives the following theorem

Theorem 28 *The* OINT *operator is true of the following statement*

- $\not\models \text{OINT}(\varphi \wedge \psi) \wedge (\varphi \not\Leftrightarrow \psi) \Leftrightarrow \text{OINT}(\varphi) \wedge \text{OINT}(\psi)$

In order to account for the revision functions the following definition is given

Definition 33 *For each world ω and time t the belief revision function BR_t^ω is a mapping from the set of belief-accessible worlds at t to the set of belief-accessible worlds at the next instant t'. Formally, $BR_t^\omega \Rightarrow B_{t'}^\omega$.*

The other revision functions can be interpreted similarly. The intuition here is that the belief revision process maps each old belief world into a corresponding new belief world. The propositions that hold in that new belief world may be quite different from those that were held in the previous worlds but no new belief worlds are introduced nor old ones deleted. For instance if we consider Figure 5.1 B_1^ω, G_1^ω, and I_1^ω represents the belief, goal and intention worlds with a branching structure. The set of belief-accessible

worlds at world ω and time 1 has a total 1-1 mapping to its corresponding goal-accessible world (denoted by f_1^ω) and intention-accessible worlds (denoted by h_1^ω. The belief revision function (BR) maps each world in B_1^ω to its corresponding world in B_2^ω and similarly for the goal (GR) and intention revision (IR) functions.

If we consider Figure 5.1 the belief world at time point 1 has three paths each ending with the propositions $(p, q), (p)$ and (q) respectively. In the new belief world at time point 2 only one path leading to (q) is left. This is because belief revision has resulted in all paths leading to (p) as a result of which (p) is removed. In the case of intention-accessible world which is a sub world of belief-accessible world at time point 1 there are two paths leading to (q). This world is filtered through the belief-accessible world at time point 2 resulting in an intention-accessible world with just one path ending with the proposition (q). This shows how the intention revision function (IR_1^ω) is related to the belief revision function and the previous intentions of the agent. The same process could be defined for all the belief-accessible worlds (and their corresponding intention-accessible worlds) of an agent at any given time point. G&R gives the following definition for belief-filtration

$$\forall b \in B_t^\omega,\ paths(IR_t^\omega(h_t^\omega(b))) \ = \ paths(BR_t^\omega(b)) \ \cap \ paths(h_t^\omega(b)) \ \neq \emptyset.$$

which states that for all belief-accessible worlds at world ω and time t, the set of paths of each revised intention-accessible world corresponding to each belief-accessible world are equal to the intersection of the paths of the revised belief-accessible world and the previous intention-accessible world. In order to ensure that an agent's new intentions are compatible with its new goals the intention-accessible worlds are filtered through the corresponding revised goal-accessible worlds to obtain new intention-accessible worlds. Hence the new intention paths is an intersection of new believed paths, new goal paths and old intention paths. Formally,

$$\forall b \in B_t^\omega,\ paths(IR_t^\omega(h_t^\omega(b))) \ = \ paths(BR_t^\omega(b)) \ \cap \ paths(DR_t^\omega(f_t^\omega(b))) \cap \\ paths(h_t^\omega(b)) \ \neq \emptyset$$

Based on these constraints G&R gives a set of axioms which together with the basic BDI axioms forms what they call a *dynamic* BDI system.

It should be noted that G&R's formalism as outlined above is a semantic approach to intention reasoning as opposed to the syntactic approach of K&P given in the previous section. Though the OINT modality of G&R seems to have close connection with that of *scenario* of K&P the former is based on a branching temporal model whereas the later have a static base. Moreover G&R's formalism deals with intention revision, belief revision etc.

which is not addressed in K&P. Other important theories of intention we would like to mention in this regard is that of [25, 116, 117, 123, 23]. Most of these theories aim to capture the dynamics involved in intention reasoning from a modal logic point of view with the help of additional structures. In [23] there is no discussion as to the dynamic behavior involved in intention reasoning but they outline a logic that provides a formal specification and decidable inference mechanism of intention consequences. We do not want to outline these theories as our aim is to account for a non-monotonic theory of intention.

5.3 The Case for Non-Monotonic Reasoning

An important classification of intention that is useful in computer science is that of *intending* versus *doing intentionally* [98], where the former involves the true intentions or preferences of the agent whereas the latter applies to the actions or states that the agent performs or brings about but not with any prior intention to do so. Based on this division Bratman classifies intentions as *deliberative, non-deliberative and policy-based*. When an agent i has an intention of the form $\mathrm{INT}_i^{t_1}\varphi, t_2$ (read as *agent i intends at t_1 to φ at t_2*) as a process of *present* deliberation, then it is called *deliberative intention*. On the other hand if the agent comes to have such an intention not on the basis of present deliberation, but at some earlier time t_0 and have retained it from t_0 to t_1 without reconsidering it then it is called *non-deliberative*. There can be a third case when intentions can be general and concern potentially recurring circumstances in an agent's life. Such general intentions constitute *policy-based intentions*, and is defined as follows:

Definition 34 *An agent i has a general-(policy/intention) to φ in circumstances of type ψ and i notes at t_1 that i am (will be) in a ψ-type circumstance at t_2, and thereby arrive at an intention to φ at t_2.*

The difference here is that there is no *present* deliberation concerning the action to be performed as the agent already has a general intention to do a particular action (*doing intentionally*). Whether the agent is able to perform that action or not depends on the circumstances.

When dealing with such general policies/intentions (hereafter intentions), we have to take into account two cases. General intentions could be either (1) *periodic* or (2) *circumstance-triggered*. They are *periodic* in the sense that their occasion for execution is guaranteed by the mere passage

of a specific interval of time. For instance, the general intention of patching up and rebooting the Unix server, *hobbit* in our department on every friday at 7pm. In contrast to this, general intention could be *circumstance triggered* as in the case of being *Super-User if one is Root*. Its occasion is not guaranteed by the mere passage of time but require that certain specific circumstances obtain. In both cases one can find that the general intention has an underlying defeasible nature. The defeasible nature is explained as follows. Consider the above example for *circumstance-triggered* general intention:

$$SU(X) \Rightarrow Root(X) \tag{5.6}$$

which means, (super-users are typically root). Suppose, there exists an agent i (a software program) that monitors tasks related to giving root permissions as and according to whether a user is a normal-user (NU) or Super-User (SU) and i has a general intention like 5.6. This general intention has a defeasible nature in the sense that, if i knows that X is a SU then i may conclude that X is Root, unless there is other evidence suggesting that X may not be root (for instance, when X has only read and write permissions but not execute permission). But this does not mean that the agent i should know all such conditions but, only those he considers necessary to the intended outcome and that he/she isn't confident of their being satisfied. Hence our definition of general intention boils down to:

> *An agent intends all the necessary consequences of his performing his general intention and he isn't confident of their being satisfied.*

In order to intend the necessary consequence the agent has to make sure that all the evidence to the contrary has been defeated which basically is a defeasible logic conclusion. This is different from the usual NML interpretation where the agent intends all the consequences.

The formation of such general policies helps in extending the influence of deliberation as it is a partial solution to the problems posed by our limited resources for calculation and deliberation at the time of action. General policies also facilitate co-ordination. It may sometimes be easier to appreciate expectable consequences (both good and bad) of general ways of acting in recurrent circumstances than to appreciate the expectable consequences of a single case.

5.4 Logical Omniscience and Non-Monotonicity

As we mentioned before, most of the theories based on NMLs interpret intention as a unary modal operator in Kripkean semantics which makes it vulnerable to the problem of *Logical-Omniscience*. The problem in its general form as stated in [123] is as follows: (where X could represent a mental state, for example INT.

1. $\models \mathbf{X}\varphi \wedge \mathbf{X}(\varphi \rightarrow \psi) \Rightarrow \mathbf{X}\psi$ (*side-effect problem*)
2. $\models \varphi \rightarrow \psi \Rightarrow \models \mathbf{X}\varphi \rightarrow \mathbf{X}\psi$ (*side-effect problem*)
3. $\models \varphi \Leftrightarrow \psi \Rightarrow \models \mathbf{X}\varphi \Leftrightarrow \mathbf{X}\psi$ (*side-effect problem*)
4. $\models \varphi \Rightarrow \models \mathbf{X}\varphi$ (*transference-problem*)
5. $\models (\mathbf{X}\varphi \wedge \mathbf{X}\psi) \rightarrow \mathbf{X}(\varphi \wedge \psi)$ (*unrestricted combining*)
6. $\models \mathbf{X}\varphi \rightarrow \mathbf{X}(\varphi \vee \psi)$ (*unrestricted weakening*)
7. $\models \neg(\mathbf{X}\varphi \wedge \mathbf{X}\neg\varphi)$

None of these properties except for (7) is valid when we take intention into consideration. For instance, consider a situation where an agent i goes to the bookstore with the intention of buying a *paper-back* and also with the intention of buying a *magazine* because he has a general intention to buy them.[1] Hence according to (5) it could be formally given as:

$$\mathrm{INT}_i(\mathit{paperback}) \wedge \mathrm{INT}_i(\mathit{magazine}) \rightarrow \mathrm{INT}_i(\mathit{paperback} \wedge \mathit{magazine})$$

But this general intention is *defeasible* in the sense that at the bookstore the agent might find that he doesn't have enough money to buy both of them and hence drops intention to buy each of them and now only intends to buy one of them. NMLs fail to account for such type of reasoning. In Sugimoto [120] an extra notion of *preference* is added and an ordering among the preferences is introduced to capture the desired effect. But we argue that, in general, such intentions are defeasible and hence a non-monotonic reasoning system would be more efficient for such occasions. The above example could be stated in a non-monotonic setup as

(1) *paper-back* (X) $\Rightarrow$ *buy*(X)
(2) *magazine* (X) $\Rightarrow$ *buy*(X)
(3) *costly* (X) $\rightsquigarrow \neg$*buy*(X)

[1] The example is a slightly modified one as given in [120].

where (1) and (2) are premises which reflects the agents general intention of buying a paper-back and magazine unless there is other evidence like (3) suggesting that he/she may not be able to buy (the meaning of the different types of arrows will be explained in the next section). When intention is formalised in the background of NMLs it is often the case that the agent has to have a complete description of the environment before-hand or has to be omniscient in the sense of knowing all the consequences. Classically the logical omniscience problem amounts to say that an agent has to compute all consequences of its own theory. It is obvious that some of the consequences are not intended as shown above. Moreover in classical NML the set of consequences is infinite. Hence we need a system like DL (defeasible logic) which is easily implementable and where the set of consequences consists of the set of literals occurring in the agent theory i.e. in the knowledge base, which is finite.

5.5 Overview of Defeasible Logic

As shown in the previous section, reasoning about general intention has a defeasible nature (in the sense that consequences may be overridden) and hence we need an efficient and easily implementable system to capture the required defeasible instances. Defeasible logic, as developed by Nute [96, 95] with a particular concern about computational efficiency and developed over the years by [9, 4, 3] is our choice. The reason being ease of implementation [88], flexibility [3] (it has a constructively defined and easy to use proof theory) and it is efficient: It is possible to compute the complete set of consequences of a given theory in linear time [87]. We do not address any semantic issues in this paper but the *argumentation semantics* given in [58] could be straightforwardly extended to the present case.

We begin by presenting the basic ingredients of DL. A defeasible theory contains five different kinds of knowledge: facts, strict rules, defeasible rules, defeaters, and a superiority relation. We consider only essentially propositional rules. Rules containing free variables are interpreted as the set of their variable-free instances.

Facts are indisputable statements, for example, "Vineet is a System Administrator". In the logic, this might be expressed as $SA(vineet)$.

Strict rules are rules in the classical sense: whenever the premises are indisputable (e.g., facts) then so is the conclusion. An example of a strict rule is "System-Administrators are Super-Users". Written formally:

$$SA(X) \rightarrow SU(X).$$

Defeasible rules are rules that can be defeated by contrary evidence. An example of such a rule is "Super-Users are typically root"; written formally:

$$SU(X) \Rightarrow Root(X).$$

The idea is that if we know that someone is a super-user, then we may conclude that he/she is root, *unless there is other evidence suggesting that he/she may not be root.*

Defeaters are rules that cannot be used to draw any conclusions. Their only use is to prevent some conclusions. In other words, they are used to defeat some defeasible rules by producing evidence to the contrary. An example is "If a user is normal-user then he might not be a root". Formally:

$$NU(X) \rightsquigarrow \neg Root(X).$$

The main point is that the information that a user is NU is not sufficient evidence to conclude that he/she is not root. It is only evidence that the user *may* not be able to become root. In other words, we don't wish to conclude $\neg Root$ if NU, we simply want to prevent a conclusion $Root$.

The *superiority relation* among rules is used to define priorities among rules, that is, where one rule may override the conclusion of another rule. For example, given the defeasible rules

$$\begin{array}{lrl} r: & SU & \Rightarrow Root \\ r': & RW & \Rightarrow \neg Root \end{array}$$

which contradict one another, no conclusive decision can be made about whether a Super-User with a *read, write* (RW) permission can be root. But if we introduce a superiority relation $>$ with $r' > r$, then we can indeed conclude that the Super-User cannot be root. The superiority relation is required to be acyclic. It turns out that we only need to define the superiority relation over rules with contradictory conclusions.

A *rule* r consists of its *antecedent* (or *body*) $A(r)$ ($A(r)$ may be omitted if it is the empty set) which is a finite set of literals, an arrow, and its *consequent* (or *head*) $C(r)$ which is a literal. Given a set R of rules, we denote the set of all strict rules in R by R_s, the set of strict and defeasible rules in R by R_{sd}, the set of defeasible rules in R by R_d, and the set of defeaters in R by R_{dft}. $R[q]$ denotes the set of rules in R with consequent q. If q is a literal, $\sim q$ denotes the complementary literal (if q is a positive literal p then $\sim q$ is $\neg p$; and if q is $\neg p$, then $\sim q$ is p).

A *defeasible theory* D is a triple $(F, R, >)$ where F is a finite set of facts, R a finite set of rules, and $>$ a superiority relation on R.

A *conclusion* of D is a tagged literal and can have one of the following four forms:

$+\Delta q$, which is intended to mean that q is definitely provable in D (i.e., using only facts and strict rules).

$-\Delta q$, which is intended to mean that we have proved that q is not definitely provable in D.

$+\partial q$, which is intended to mean that q is defeasibly provable in D.

$-\partial q$ which is intended to mean that we have proved that q is not defeasibly provable in D.

Provability is based on the concept of a *derivation* (or proof) in $D = (F, R, >)$. A derivation is a finite sequence $P = (P(1), \ldots P(n))$ of tagged literals satisfying four conditions (which correspond to inference rules for each of the four kinds of conclusion). $P(1..i)$ denotes the initial part of the sequence P of length i

$+\Delta$: If $P(i+1) = +\Delta q$ then
(1) $q \in F$ or
(2) $\exists r \in R_s[q]\ \forall a \in A(r) : +\Delta a \in P(1..i)$

$-\Delta$: If $P(i+1) = -\Delta q$ then
(1) $q \notin F$ and
(2) $\forall r \in R_s[q]\ \exists a \in A(r) : -\Delta a \in P(1..i)$

The definition of Δ describes just forward chaining of strict rules. For a literal q to be definitely provable we need to find a strict rule with head q, of which all antecedents have been definitely proved previously. And to establish that q cannot be proven definitely we must establish that for every strict rule with head q there is at least one antecedent which has been shown to be non-provable.

$+\partial$: If $P(i+1) = +\partial q$ then either
(1) $+\Delta q \in P(1..i)$ or
(2) (2.1) $\exists r \in R_{sd}[q] \forall a \in A(r) : +\partial a \in P(1..i)$ and
(2.2) $-\Delta{\sim}q \in P(1..i)$ and
(2.3) $\forall s \in R[{\sim}q]$ either
(2.3.1) $\exists a \in A(s) : -\partial a \in P(1..i)$ or
(2.3.2) $\exists t \in R_{sd}[q]$ such that
$\forall a \in A(t) : +\partial a \in P(1..i)$ and $t > s$

Let us work through this condition. To show that q is provable defeasibly we have two choices: (1) We show that q is already definitely provable; or (2) we need to argue using the defeasible part of D as well. In particular, we require that there must be a strict or defeasible rule with head q which can be applied (2.1). But now we need to consider possible "attacks", that is, reasoning chains in support of $\sim q$. To be more specific: to prove q defeasibly we must show that $\sim q$ is not definitely provable (2.2). Also (2.3) we must consider the set of all rules which are not known to be inapplicable and which have head $\sim q$ (note that here we consider defeaters, too, whereas they could not be used to support the conclusion q; this is in line with the motivation of defeaters given earlier). Essentially each such rule s attacks the conclusion q. For q to be provable, each such rule s must be counterattacked by a rule t with head q with the following properties: (i) t must be applicable at this point, and (ii) t must be stronger than s. Thus each attack on the conclusion q must be counterattacked by a stronger rule. In other words, r and the rules t form a team (for q) that defeats the rules s. In an analogous manner we can define $-\partial q$ as

$-\partial$: If $P(i+1) = -\partial q$ then
(1) $-\Delta q \in P(1..i)$ and
(2) (2.1) $\forall r \in R_{sd}[q]\ \exists a \in A(r) : -\partial a \in P(1..i)$ or
(2.2) $+\Delta{\sim}q \in P(1..i)$ or
(2.3) $\exists s \in R[\sim q]$ such that
(2.3.1) $\forall a \in A(s) : +\partial a \in P(1..i)$ and
(2.3.2) $\forall t \in R_{sd}[q]$ either
$\exists a \in A(t) : -\partial a \in P(1..i)$ or $t \not> s$.

The purpose of the $-\partial$ inference rules is to establish that it is not possible to prove $+\partial$. This rule is defined in such a way that all the possibilities for proving $+\partial q$ (for example) are explored and shown to fail before $-\partial q$ can be concluded. Thus conclusions tagged with $-\partial$ are the outcome of a constructive proof that the corresponding positive conclusion cannot be obtained.

Sometimes all we want to know is whether a literal is *supported*, that is if there is a chain of reasoning that would lead to a conclusion in absence of conflicts. This notion is captured by the following proof conditions:

$+\Sigma$: if $P(i+1) = +\Sigma p$ then
(1) $+\Delta p \in P(1..i)$ or
(2) $\exists r_{sd}[p] : \forall a \in A(r) + \Sigma a \in P(1..i)$.

and

$-\Sigma$: if $P(i+1) = -\Sigma p$ then
(1) $-\Delta p \in P(1..i)$ and
(2) $\forall r_{sd}[p] \exists a \in A(r) : -\Sigma a \in P(1.i)$

The notion of support corresponds to monotonic proofs using both the monotonic (strict rules) and non-monotonic (defeasible rules) parts of defeasible theories.

5.6 Defeasible Logic for Intentions

As we have seen in section 5.4 NMLs have been put forward to capture the intensional nature of mental attitudes such as, for example, intention. Usually modal logics are extensions of classical propositional logic with some intensional operators. Thus any classical (normal) modal logic should account for two components: (1) the underlying logical structure of the propositional base and (2) the logic behavior of the modal operators. Alas, as is well-known, classical propositional logic is not well suited to deal with real life scenarios. The main reason is that the descriptions of real-life cases are, very often, partial and somewhat unreliable. In such circumstances classical propositional logic might produce counterintuitive results insofar as it requires complete, consistent and reliable information. Hence any modal logic based on classical propositional logic is doomed to suffer from the same problems.

On the other hand the logic should specify how modalities can be introduced and manipulated. The common rules for modalities are

$$\frac{\vdash \varphi}{\vdash \Box\varphi}\text{Necessitation}$$

$$\frac{\vdash \varphi \supset \psi}{\vdash \Box\varphi \supset \Box\psi}\text{RM}$$

Consider the necessitation rule of normal modal logic which dictates the condition that an agent knows all the valid formulas and thereby all the tautologies. Such a formalisation might suit for the knowledge an agent has but definitely not for the intention part. Moreover an agent need not be intending all the consequences of a particular action it does. It might be the case that it is not confident of them being successful. Thus the two rules are not appropriate for a logic of intention.

A logic of policy-based intention should take care of the underlying principles governing such intentions. It should have a notion of the direct and indirect knowledge of the agent, where the former relates to facts as

literals whereas the latter to that of the agent's theory of the world in the form of rules. Similarly the logic should also be able to account for general intentions as well as the policy-based (derived ones) intentions of the agent.

Accordingly a defeasible intention theory is a structure $(F, R^K, R^I, >)$ where, as usual F is a set of facts, R^K is a set of rules for knowledge (i.e., $\rightarrow_K$, $\Rightarrow_K$, $\rightsquigarrow_K$), R^I is a set of rules for intention (i.e., $\rightarrow_I$, $\Rightarrow_I$, $\rightsquigarrow_I$), and $>$, the superiority relation, is a binary relation over the set of rules (i.e., $> \subseteq (R^K \cup R^I)^2$).

Intuitively, given an agent, F consists of the information the agent has about the world and its immediate intentions; R^K corresponds to the agent's theory of the world, while R^I encodes its policy and $>$ its strategy (or its preferences). The policy part of a defeasible theory captures both intentions and goals. The main difference is the way the agent perceives them: goals are possible outcomes of a given context while intentions are the actual goals the agent tries to achieve in the actual situation. In other words goals are the choices an agent has and intentions are the chosen goals; in case of conflicting goals (policies) the agent has to evaluate the pros and cons and then decide according to its aims (preferences), which are encoded by the superiority relation (the agent's strategy).

In what follows we provide the appropriate inference rules for intentions, and we identify strong intentions – i.e., intentions for which there are no alternatives – using $\pm\Delta_I$; goals using $\pm\Sigma_I$, and intentions using $\pm\partial_I$.

In order to correctly capture the notion of intention we extend the signature of the logic with the modal operator INT; thus if l is literal then INTl and $\neg$INTl are modal literals. However we impose some restrictions on the form of the rules: modal literals can only occur in the antecedents of rules for intention.

Derivability for knowledge ($\pm\Delta_K$, $\pm\partial_K$) has the same conditions as those given for derivability in Section 5.5. It is true that the complete and accurate definition of the inference conditions is cumbersome but the intuition is natural and easy to understand. The conditions for deriving an intention are as follows:

$+\Delta_I$: if $P(i+1) = +\Delta_I p$ then
(1) INT$p \in F$ or
(2) $\exists r \in R_s^K[p] \forall a \in A(r) : +\Delta_I a \in P(1..i)$ or
(3) $\exists r \in R_s^I[p]$ such that
(3.1) $\forall$INT$a \in A(r) : +\Delta_I a \in P(1..i)$ and
(3.2) $\forall a \in A(r) : +\Delta_K a \in P(1..i)$.

To prove a strong intention, we need either that the intention is unconditional (1), i.e., a basic intention, or that we have a strict rule for intention

(an irrevocable policy) whose antecedent is indisputable (3). However we have another case (2): if an agent knows that B is an indisputable consequence of A, and the agent strongly intends A, then it must intend B. This is in contrast with the NML interpretation whereby the agent has to intend all the consequences of his/her intention.

$-\Delta_I$: if $P(i+1) = -\Delta_I p$ then
(1) INT$p \notin F$ and
(2) $\forall r \in R_s^K[p]$
(2.1) $\exists a \in A(r) : -\Delta_K a \in P(1..i)$ or
(2.2) $\exists a \in A(r) : -\Delta_I a \in P(1..i)$; and
(3) $\forall r \in R_s^I[p]$ either
(3.1) $\exists$INT$a \in A(r) : -\Delta_I a \in P(1..i)$ or
(3.2) $\exists a \in A(r) : -\Delta_K a \in P(1..i)$.

To prove that a strong intention A does not hold ($-\Delta_I A$), first, A should not be a basic intention (1); then we have to discard all possible reasons in favor of it. If A is a definite consequence of B, that is $B \rightarrow_K A \in R^K$, we can disprove it if we can show that (2.1) B is not the case (i.e., $-\Delta_K B$) or (2.2) B is not strongly intended (i.e., $-\Delta_I B$). In case of strict policies for A (3), such as, for example the strict rule for intention INT$B, C \rightarrow_I A$, we have to show that either B is not strongly intended (3.1), or the fact triggering the policy is not the case (3.2).

At the other extreme we have goals: literals supported by evidence and basic intentions.

$+\Sigma_I$: if $P(i+1) = +\Sigma_I p$ then
(1) INT$p \in F$ or
(2) $\exists r \in R_{sd}^K[p] \forall a \in A(r) : +\Sigma_I a \in P(1..i)$ or
(3) $\exists r \in R_{sd}^I[p]$ such that
(3.1) $\forall$INT$a \in A(r) : +\Sigma_I a \in P(1..i)$ and
(3.2) $\forall a \in A(r) : +\Sigma_K a \in P(1..i)$.

$-\Sigma_I$: if $P(i+1) = -\Sigma_I p$ then
(1) INT$p \notin F$ and
(2) $\forall r \in R_{sd}^K[p]$
(2.1) $\exists a \in A(r) : -\Sigma_K a \in P(1..i)$ or
(2.2) $\exists a \in A(r) : -\Sigma_I a \in P(1..i)$; and
(3) $\forall r \in R_{sd}^I[p]$ either
(3.1) $\exists$INT$a \in A(r) : -\Sigma_I a \in P(1..i)$ or
(3.2) $\exists a \in A(r) : -\Sigma_K a \in P(1..i)$.

The inference conditions for goals are very similar to those for strong intentions; essentially they are monotonic proofs using both the monotonic part (strict rules) and the supportive non-monotonic part (defeasible rules) of a defeasible theory.

On the other hand to capture intentions we have to use the superiority relations to resolve conflicts. Thus we can give the following definition for the inference rules for $\pm\partial_I$.

$+\partial_I$: if $P(i+1) = +\partial_I p$ then
(1) $+\Delta_I p \in P(1..i)$ or
(2) (2.1) $-\Delta_K{\sim}p, -\Delta_I{\sim}p \in P(1..i)$ and
(2.2) either
(2.2.1) $\exists r \in R^K_{sd}[p] \forall a \in A(r) : +\partial_I a \in P(1..i)$, or
(2.2.2) $\exists r \in R^I_{sd}[p]$ $\forall \mathrm{INT} a \in A(s) : +\partial_I a \in P(1..i)$ and
$\forall a \in A(s) : +\partial_K a \in P(1..i)$; and
(2.3) $\forall s \in R[{\sim}p]$ either
(2.3.1) if $s \in R^K[{\sim}p]$ then
$\exists a \in A(s) : -\partial_I a \in P(1..i)$ and
$\exists b \in A(s) : -\partial_K b \in P(1..i)$; and
if $s \in R^I[{\sim}p]$ then either
$\exists \mathrm{INT} a \in A(s) : -\partial_I a \in P(1..i)$ or
$\exists a \in A(s) : -\partial_K a \in P(1..i)$; or
(2.3.2) $\exists t \in R[p]$ such that $t > s$ and
if $t \in R^K[p]$ then $\forall a \in A(t) : +\partial_K a$ or
$\forall a \in A(t) : +\partial_I a$; and
if $t \in R^I[p]$ then $\forall a \in A(t) : +\partial_K a$ and
$\forall \mathrm{INT} a \in A(t) : +\partial_I a$.

The conditions for proving defeasible intentions are essentially the same as those given for defeasible derivations in Section 5.5. The only difference is that at each stage we have to check for two cases, namely: (1) the rule used is a rule for an intention; (2) the rule is a rule for knowledge. In the first case we have to verify that factual antecedent are defeasibly proved/disproved using knowledge ($\pm\partial_K$), and intentional antecedent are defeasibly proved/disproved using intention ($\pm\partial_I$). In the second case we have to remember that a conclusion of a factual rule can be transformed in an intention if all the literals in the antecedent are defeasibly intended.

$-\partial_I$: if $P(i+1) = -\partial_I p$ then
(1) $-\Delta_I p \in P(1..i)$ and
(2) (2.1) $+\Delta_K{\sim}p$ or $+\Delta_I{\sim}p \in P(1..i)$ or
(2.2) both

$$
\begin{array}{rr}
(2.2.1)\ \forall r \in R^K_{sd}[p] & \exists a \in A(r) : -\partial_K a \in P(1..i), \text{ and} \\
 & \exists a \in A(r) : -\partial_I a \in P(1..i); \text{ and} \\
(2.2.2)\ \forall r \in R^I_{sd}[p] & \exists \mathrm{INT} a \in A(s) : -\partial_I a \in P(1..i) \text{ or} \\
 & \exists a \in A(s) : -\partial_K a \in P(1..i); \text{ or} \\
(2.3)\ (2.3.1)\ \exists s \in R^K[\sim p] & \forall a \in A(s) : +\partial_K a \text{ or} \\
 & \forall a \in A(s) : +\partial_I a, \text{ or} \\
\exists s \in R^K[\sim p] & \forall a \in A(s) : +\partial_K a \text{ and} \\
 & \forall \mathrm{INT} a \in A(s) : +\partial_I a; \text{ and} \\
(2.3.2)\ \exists t \in R[p] & \text{such that } t > s \text{ and} \\
\text{if } t \in R^K[p] & \text{then } \forall a \in A(t) : +\partial_K a \text{ or} \\
 & \forall a \in A(t) : +\partial_I a; \text{ and} \\
\text{if } t \in R^I[p] & \text{then } \forall a \in A(t) : +\partial_K a \text{ or} \\
 & \forall \mathrm{INT} a \in A(t) : +\partial_I a.
\end{array}
$$

The intuition behind the definition of $-\partial_I$ is a combination of the motivation for $-\partial$ and the intuition of $-\Delta_I$.

We want to illustrate some of the aspects of derivability by means of examples. For instance, consider the following scenario: If it does not rain we intend to play cricket, and if we intend to play cricket we intend to stay outdoor. This example can be formalized as follows

$$
\begin{array}{c}
\neg \mathit{rain} \Rightarrow_I \mathit{cricket} \\
\mathrm{INT}\mathit{cricket} \Rightarrow_I \mathit{outdoor}
\end{array}
$$

Once the fact $\neg \mathit{rain}$ is supplied we can derive $+\partial_I \mathit{cricket}$, and then the intention of staying outdoor ($+\partial_I \mathit{outdoor}$). However the same intention cannot be derived if the fact *cricket* is given.

If Vineet intend to travel to Italy then he intend to travel to Europe since Italy is in Europe. This argument can be formalized as

$$\mathit{Italy} \rightarrow_K \mathit{Europe}$$

plus the basic intention $\mathrm{INT}\mathit{Italy}$. The conclusion $+\Delta_I \mathit{Europe}$ follows from clause (2) of $+\Delta_I$.

Most of the BDI systems are able to express positive and negative introspection of belief and intentions. Those notions are encoded, respectively, by the following axioms.

$$
\begin{array}{c}
\mathrm{INT}\phi \rightarrow \mathrm{BEL}(\mathrm{INT}\phi) \\
\neg\mathrm{INT}\phi \rightarrow \mathrm{BEL}(\neg\mathrm{INT}\phi)
\end{array}
$$

One of the main effect of positive (resp. negative) introspection is the ability of using established (resp. rejected) intentions in epistemic contexts to

derive (resp. prevent the derivation of) other intentions. But this is what is done in Clause 2 of $+\Delta_I$, Clause 2.2.1 of $+\partial_I$, for positive introspection, and Clause 2.2 of $-\Delta_I$ and Clause 2.2.1 of $-\partial_I$ for negative introspection.

The purpose of the $-\Delta$ and $-\partial$ inference rules is to establish that it is not possible to prove a corresponding tagged literal. These rules are defined in such a way that all the possibilities for proving $+\partial p$ (for example) are explored and shown to fail before $-\partial p$ can be concluded. Thus conclusions with these tags are the outcome of a constructive proof that the corresponding positive conclusion cannot be obtained.

As a result, there is a close relationship between the inference rules for $+\partial$ and $-\partial$, (and also between those for $+\Delta$ and $-\Delta$, and $+\Sigma$ and $-\Sigma$). The structure of the inference rules is the same, but the conditions are negated in some sense. This feature allows us to prove some properties showing the well behaviour of defeasible logic.

Theorem 29 *Let $\# = \Delta_K, \partial_K, \Sigma_K, \Delta_I, \partial_I, \Sigma_I$, and D be a defeasible theory. There is no literal p such that $D \vdash +\#p$ and $D \vdash -\#p$.*

Proof. This Theorem is an immediate consequence of the *Principle of strong Negation* and Theorem 2 of [3]. Indeed it is straightforward to prove that the conditions for $+\#$ are the strong negation of those for $-\#$, and vice-versa. □

Intuitively the above theorem states that no literal is simultaneously provable and demonstrably unprovable, thus it establishes the coherence of the defeasible logic presented in this chapter.

Theorem 30 *Let D be a defeasible theory, and $M \in \{K, I\}$. $D \vdash +\partial_M p$ and $D \vdash +\partial_M \sim p$ iff $D \vdash +\Delta_M p$ and $D \vdash +\Delta_M \sim p$.*

Proof. For $M = K$ see Proposition 3.3 in [4]. □

This theorem gives the consistency of defeasible logic. In particular it affirms that it is not possible to obtain conflicting intentions ($+\partial_I p$ and $+\partial_I \sim p$) unless the information given about the environment is itself inconsistent. Notice, however, that the theorem does not cover goals (Σ_I). Indeed, it is possible to have conflicting goals.

Let D be a defeasible theory. With Δ_K^+ we denote the set of literals strictly provable using the epistemic (knowledge) part of D, i.e., $\Delta_K^+ = \{p : D \vdash +\Delta_K p\}$. Similarly for the other proof tags.

Theorem 31 *For every defeasible theory D, and $M \in \{K, I\}$*

1. $\Delta^+_M \subseteq \partial^+_M \subseteq \Sigma^+_M$;

2. $\Sigma^-_M \subseteq \partial^-_M \subseteq \Delta^-_M$.

Proof. We prove 1, and 2 is a consequence of 1 and the principle of strong negation [3]. For $M = K$ see [4], since the conditions for knowledge are those for derivability in DL. The inclusion $+\Delta_I \subseteq +\partial_I$ is immediate from condition 1 of $+\partial_I$. For the other inclusion, i.e., $+\partial_I \subseteq +\Sigma_I$ we notice that if we restrict ourselves to strict rules in $+\Sigma_I$ we obtain clause 1 of $+\partial_I$, the basic case of the recursive definiton of derivation in DL. Moreover clause 2 and 3 of $+\Sigma_I$ correspond to clause 2.2.2 and 2.2.3 of $+\partial_I$. □

This theorem states that strict intentions are intentions ($\Delta^+_I \subseteq \partial^+_I$), and intentions are goals ($\partial^+_I \subseteq \Sigma^+_I$), which corresponds to the well-known BDI principle

$$\mathrm{INT}\phi \rightarrow \mathrm{GOAL}\phi.$$

At the same time, we have that $\Delta^+_K \subseteq \partial^+_K$. Thus if we assume that Δ_K corresponds to knowledge and ∂_K corresponds to belief we obtain

$$\mathrm{KNOW}\phi \rightarrow \mathrm{BEL}\phi,$$

the standard BDI axiom relating the two epistemic notions.

The proposed theory of intention satisfies many of the properties outlined by Bratman in [12]. The role of intention as a *conduct-controlling* pro-attitude rather than *conduct-influencing* is clearly illustrated in the elaborate proof-theory outlined for the types of intention. The proposed theory supports the fact that the rationality of an agent for his intention depends on the rationality of the relevant processes leading to that intention where the relevant processes includes using superiority relations to resolve conflicts as well as satisfying the rules of inclusion as shown in Theorem 30. The new approach provides a good formalisation as to the relation between *guiding intention* and *intentional action* termed as *historical principle of policy-based rationality* in [12]. The problem in general is to account for the rationality of an agent in performing a particular policy-based intention from a general policy. In our approach the defeasibility of general policies makes it possible to block/not block the application of the policy to the particular case without abandoning the policy. This idea is explained further in the next section.

5.7 Revisiting Intention Reconsideration and Commitment Strategies

In the preceding section we showed how to account for a defeasible logic of intention. This section tries to relate the new theory with an intensional system like BDI and tries to answer questions related to *intention-reconsideration* and *commitment* in such systems. As noted earlier we are concerned only with the intention part and we assume the existence of factual knowledge. One of the issues in the design of agents that are based on the models of intention is that of when to reconsider intentions. An agent cannot simply maintain an intention, once adopted, without ever stopping to reconsider. It is necessary from time to time for an agent to check whether the intention has been achieved or whether it is no longer achievable. All the existing models of intentional systems are based on deliberative or non-deliberative theory of intention where reconsideration of intention is hard as the deliberation time required for such an act is huge. Whereas in the case of our policy-based intention framework the defeasibility of general intentions makes it possible to block the application of the intention to the particular case without abandoning/reconsidering the intention. Consider the following scenario:

Example 3 *Suppose that the software program i, as mentioned in the first section, uses a* policy server *to grant root access to its users. One of the rule the program uses corresponds to* $SU(X) \Rightarrow Root(X)$. *In-order for the program to successfully grant permission or to detect a conflict, the following conditions should be satisfied according to our new approach*

For granting permission the program should either find *vineet* (if X = vineet) in the data-base of facts or it should be able to defeasibly prove it, which in turn means has to check with all the contrary evidence (for instance whether X satisfies all the root permissions etc.) On the other hand, conflict detection is done either when the program cannot definitely prove the identity of the user or when the program cannot defeasibly prove the user's identity. This approach to conflict detection is much simpler than the ones mentioned in [119]. Moreover, in our framework we have the liberty of specifying a *superiority* relation among the rules which is absent in policy description languages like *PDL* [84]. It is also the case that the four different types of conclusion in DL helps to solve the issue as to when an intention should be dropped. An intention is dropped when it is either proved definitely ($+\Delta$) or cannot be proved defintely ($-\Delta$). The two defeasible rules ($+\partial$) and ($-\partial$) allows the agent to have sufficient deliberation upon its intention.

In [107] three different commitment strategies, *blind, single-minded and open-minded* has been defined. A blindly committed agent maintains her intentions until they succeed. This could be contrasted with the defeasible logic conclusion of *definitely proving*, where the inference is made from facts and strict rules. Hence a blindly committed agent will maintain her intention forever until he/she is able to prove it definitely ($+\Delta$). In other words as long as there is no conflict between the intentions a blindly committed agent succeeds in his/her pursuit. This could be formally represented as

$$\mathrm{INT}(\varphi) \wedge \mathrm{INT}(\psi) \rightarrow \neg conflict(\varphi, \psi)^2$$

On the other hand a *single-minded agent* maintains her intention as long as she believes her intentions to be achievable. This means that the agent maintains her intentions as long as it is definitely provable ($+\Delta$) or not definitely provable ($-\Delta$). This could be described in a rule format as

$$\begin{aligned}\mathrm{GOAL}(\varphi) \wedge \mathrm{INT}(\psi) \wedge conflict(\varphi, \psi) \wedge (> (\varphi) = \ > (\psi)) \\ \rightarrow \neg \mathrm{INT}(\varphi)\end{aligned}$$

The rule states that if an agent has a goal φ and has an intention ψ and if there is a conflict between φ and ψ and the superiority relation between φ and ψ are equal then the agent does not intend φ. Hence the agent fails in definitely proving the intention as it finds a conflicting goal. This sort of reasoning is truly in agreement with the BDI principle that the INT (intention) of an agent is always supported by its GOAL i.e. $\mathrm{INT}(\varphi) \rightarrow \mathrm{GOAL}(\varphi)$ (there should be at least one intention world that is a sub-world of the goal world). Finally, an *open-minded* agent maintains her intentions as long as these intentions are her goals which means that as long as they could be achieved through either one of the inference mechanisms. Though we have taken into account only goals and intentions similar considerations are applicable in the case of beliefs and goals as well as beliefs and intentions.

5.8 Summary and Discussion

Based on Bratman's classification of intention, we have outlined a *policy-based* theory of intention which differs from the usual NML-based approaches in the sense of having a non-monotonic nature. To capture the properties involved in such intentions we adopted *defeasible logic* as the non-monotonic reasoning mechanism due to its efficiency and ease of implementation as well as the defeasible nature of policy-based intentions. The new

[2] A similar formalism appears in [121]

approach alleviates most of the problems related to logical-omniscience. We pointed out that some of the problems related to intention re-consideration could be easily understood through such an approach. The *commitment* strategy adopted by BDI-like agents was reviewed in the light of the new approach.

The approach outlined in this paper could be extended in at least two different directions.

The first is in alliance with the work done in [119, 84]. Here they outline a policy description language called *PDL* and use logic programs to reason about the policies. The main concern in that work is in tracing the *event* history that gives rise to an *action* history based on stable model semantics. In a similar manner our approach could be developed using the appropriate semantics (Kunen [43] or argumentation [58]) and developed from a logic programming point of view. The advantage in our approach is the use of the superiority relation ($>$) whereby we can mention a hierarchy between the rules and this is absent in other works.

The second direction in which our work could be extended is to define various rules required for constructing goals from beliefs, intentions from goals, intentions from beliefs etc. and giving a superiority relation among these rules. The recent work on BDI [121] seems to take this direction. On the other hand many new applications in emerging information technologies have advanced needs for managing relations such as authorization, trust and control among interacting agents (humans or artificial) [57, 50, 71, 18]. This necessitates new models and mechanisms for structuring and flexible management of those relations. The issues of automated management of organisations in terms of policies and trust relations in highly dynamic and decentralised environments has become the focus in recent years.

Finally, as we have alluded to many semantics have been devised for defeasible logic and can be adapted straightforwardly to the extension proposed here. The method developed in [43] gives a set-theoretic fixed-point construction for $\Delta^+, \partial^+, \ldots$, which leads to a logic programming characterization of defeasible logic. Programs corresponding to defeasible theories are sound and complete wrt Kunen semantics. The same technique is applicable in the present case with the obvious adjustments; however, it does not offer further insights on defeasible logic for BDI, because of the almost one-to-one correspondence between the inference conditions and the steps of the fixed-point construction. However semantics for defeasible BDI logic remains an interesting technical problem.

Chapter 6

Conclusions and Future Work

The whole is more than sum of its parts.

Aristotle

In this brief chapter we summarise the findings and contributions of the thesis and discuss future research directions.

6.1 Summary

The overall aim of this dissertation has been to study BDI logics in the background of combinations of modal logics. In order to address this aim we adopted a general combining technique called fibring/dovetailing that allows to produce combinations of BDI logics. In the course of the study we also addressed ways of incorporating action constructs in a BDI framework and also defined a non-monotonic theory of intention.

In more detail, Chapter 2 gives an overview of BDI logics by placing it in the context of modal logic in a multi-modal setting. Two combining techniques called fusion and fibring are discussed and it is argued that fibring is more suited to combine BDI logics than fusion.

In Chapter 3 we reconstructed the logical account of BDI in terms of *dovetailing*, which is a special case of fibring, with the help of the incestual schema $G^{a,b,c,d}$. In doing so we identified a set of interaction axioms for BDI which covers many of the existing BDI axioms and also make possible the generation of a large class of new ones. We showed that BDI is determined by the class of dovetailed models satisfying the a,b,c,d-incestuality properties. This should avoid the need for showing determination results

each time a new BDI system is considered. Further we identified conditions under which completeness transfers from the component logics to the fibred/dovetailed composition. This is done with the help of canonical model structures where we modify the usual canonical model construction in modal logic so as to account for fibred models. We also show *completeness preservation* in the case of interaction axioms of the type $\Box_1\varphi \Rightarrow \Box_2\varphi$. The proof uses an extension of canonical model method for modal logics together with certain *morphism* conditions. This is a major result when compared to other combining techniques like *fusion* which doesn't support any interaction axioms. We also show that the resulting multimodal system is non-homogeneous in the sense that all interactions involved could be either between different mental attitudes of the same agent or could involve same mental attitudes of different agents. This sort of classification is absent in most agent theories with interaction axioms. It was further observed that the our approach is not restricted to normal modal operators but could be extended to non-normal ones too.

In Chapter 4 we outlined two important philosophical theories on modal logics of action and ability. We showed that both of them like BDI fails to address the concept of *composite actions* of the type $\pi_1;\pi_2$ (read as π_1 *followed by* π_2). In order to accommodate composite actions we introduced two new operators *opportunity* (OPP) and *result* (RES). It is argued that the successful execution of composite actions depends on the result of its components. It is also observed that the new constructs fall in line with Bratman's classification of *intentional* action and *intending* an action. Further, we related the two new operators with the BDI framework as was given in Chapter 2. This led to the investigation of the close connection between the result of an action performed by a BDI agent and its *capability* of achieving that result. Based on this investigation we showed that both RES and OPP are necessary for a theory of composite actions. Further we introduced two more action constructs, *while* φ *do* π and *if* φ *then* π_1 *else* π_2, We then explained our intuition on the composite behaviour of results, opportunities and capabilities and presented a set of axioms that comply with this intuition. Using the various modalities present in the framework, we proposed a set of commitment axioms, that classifies a BDI agent as *blindly committed, single minded* and *open minded*.

Chapter 5 investigated defeasible logic and proposed a non-monotonic theory of intention. We started by arguing that normal modal logics was not suitable to modal intention as it often leads to the problem of *logical omniscience*. Two seminal theories on intention maintenance were reviewed and conclusion made that the intention component of a BDI agent need not always have a monotonic setup. We outlined the classification of intentions

as given by Bratman and came to the conclusion that a non-monotonic theory of intention is needed to account for *policy-based* intentions. To this end we adopted defeasible logic and proposed a defeasible theory of intention. Certain restrictions were made in the case of BDI modalities as for instance we consider a knowledge operator in the place of belief. Further, we outlined a proof theory which shows how our approach helps in the maintenance of intention-consistency in agent systems like BDI. We give three theorems of which the first two are related to properties like coherence and consistency in defeasible logic. The third theorem accounts for the interaction axioms in BDI. We conclude the chapter by showing how our approach provides an adequate explanation to *intention reconsideration* as well as the *commitment axioms* in BDI systems.

6.2 Future Work

6.2.1 Modeling Multi-Agent Systems (MAS)

We foresee several directions in which the research captured in this thesis could proceed. As previously discussed, the main aim of the thesis has been to study BDI logics in the background of combining techniques like fibring/dovetailing. BDI is a multi agent system (MAS) and hence the tools and techniques developed here could easily be applied in the case of MAS theories. MAS require the specifications of a number of potentially heterogeneous agent systems, where each agent is itself a complex system arising from the combinations of different aspects (for example, mental attitudes for BDI-like agents). The main difficulty with trying to apply combining logics to MAS theories is that while the later heavily depends on interaction axioms between the various logics to be combined, the former hardly addresses this case at all. But as we have shown in Chapter 3 it is possible to add conditions on the fibring function. These conditions could encode interactions between the two classes of models that are being combined and therefore could represent interaction axioms between the two logics. MAS theories with interaction axioms will be one of the ideal test-beds for fibring and as a result MAS theories could benefit much from this. The construction of cultural bridges between combining logics and MAS theories is a useful exercise. On the one hand this could encourage MAS theorists to recognize the need of shifting the attention from the single case analysis to the general problem of the combination of mental states. Then MAS theorists could built formal models of agency simply by considering basic well understood logics and study the possible interplay of the different

components when combined together. On the other hand the problem of interaction axioms between logic fragments in a combined logic might become more central in combining logics. Results of properties transfer, even if limited to very small classes of interaction axioms, would be of extreme importance to the whole area of MAS theories. If operating under the guide lines of a well developed theory of logic combination, properties of the logics such as completeness and decidability would be inherited from the basic components. This would permit us to develop not just a general theory for MAS but to define the criteria in order to define *appropriate* specifications of the system the AI-user has in mind. To summarise, the following are some of the possible directions in which fibring could be extended to MAS theories:

- families of interactions axioms appropriate to describe the relationships among the mental attitudes both in single agent environments and in multiple agent settings.
- generalization of fibring with richer mathematical structure and the fibring of (modal) algebras.
- Identification of the constraints on fibring required to characterize some families of interaction axioms.
- inference rules and operations on labels appropriate for capturing the constraints on fibring for interaction axioms

6.2.2 Tableaux Systems

Another possible extension is the development of tableaux systems to deal with the class of normal multimodal logics determined by the axiom $G^{a,b,c,d}$. Governatori and co-workers (see for example, [55, 5, 6]) have developed a modular labelled tableaux system (called KEM) which is able to deal with logic admitting possible world semantics. Gabbay and Governatori [43, 44] have shown that KEM can be used in the context of fibring of modal logics provided that KEM tableaux systems for the component logics exist. In [7, 55, 56] tableaux for some multi-modal logics of agency with interaction axioms have been given. However, so far, only a few interaction axioms have been studied in this respect. Thus future investigation could involve 1) KEM tableaux systems for the modal logics needed to represent BDI-agents and MAS; and 2) conditions under which interaction axioms can be represented in KEM.

6.2.3 Dynamics of Multi-Agent Systems

Of the mental attitudes in BDI (belief, goal, intention), the revision of beliefs has been the subject of considerable study in both Artificial Intelligence and Philosophy (e.g., [49, 39]). The more general problems of revising complex plans and knowledge of an agent's capabilities remain open questions. The fibred/dovetailed approach will address multi-modalities of agents in a uniform manner using the combined logic. Any reasonable progress towards identifying constraints that must be satisfied by agents while revising their mental attitudes in a changing world, would be a significant contribution to the field. Rationality postulates for belief revision exist in literature: the AGM postulates by Alchourrón, Gärdenfors and Makinson [1] address the case of a static environment, while those of Katzuno and Mendelzon [72] apply in the case of a dynamic environment. However, there has been almost no work on generalizing these results to the case of modal logic, with the notably exception of [53]. In particular it is possible 1) to investigate the applicability of the method of [53] in the unified framework for multi-modal logics; and 2) to extend and modifying the existing postulates for belief change in order to provide more appropriate ground for theory change in intensional contexts, and for the dynamic of multi-agent systems.

Bibliography

[1] Carlos Alchourron, Peter Gardenfors, and David MAkinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 185.

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

[3] Grigoris Antoniou, David Billington, Guido Governatori, and Michael Maher. A flexible framework for defeasible logics. In *Proc. American National Conference on Artificial Intelligence (AAAI-2000)*, pages 401–405. AAAI/MIT Press, 2000.

[4] Grigoris Antoniou, David Billington, Guido Governatori, and Michael J. Maher. Representation results for defeasible logic. *ACM Transactions on Computational Logic*, 2(2):255–287, April 2001.

[5] Alberto Artosi, Paola Benassi, Guido Governatori, and Antonino Rotolo. Shakespearian modal logic: A labelled treatment of modal identity. In Marcus Kracht, Maarten de Rijke, Heinrich Wansing, and Michael Zakharyaschev, editors, *Advances in Modal Logic. Volume 1*, pages 1–21. CSLI Publications, Stanford, 1998.

[6] Alberto Artosi, Guido Governatori, and Antonino Rotolo. Labelled tableaux for non-monotonic reasoning: Cumulative consequence relations. *Journal of Logic and Computation*, 12(6):1037–1060, February 2002.

[7] Alberto Artosi, Guido Governatori, and Giovanni Sartor. Towards a computational treatment of deontic defeasibility. In Mark Brown and José Carmo, editors, *Deontic Logic Agency and Normative Systems*, Workshop on Computing, pages 27–46, Berlin, 1996. Springer-Verlag.

[8] Matteo Baldoni. *Normal Multimodal Logics: Automatic Deduction and Logic Programming Extension*. PhD thesis, Departimento di informatica, Universita degli Studi di Torino, Italy, 1998.

[9] David Billington. Defeasible logic is stable. *Journal of Logic and Computation*, 3:370–400, 1993.

[10] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, 2001.

[11] Rafael H. Bordini, Ana L. C. Bazzan, Rafael de O. Jannone, Daniel M. Basso, Rosa M. Vicari, and Victor R. Lessor. AgentSpeak(XL): Efficient intention selection in BDI agents via decision-theoretic task scheduling. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems, (AAMAS 2002)*. ACM, 2002.

[12] Michael E. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.

[13] R. A. Brooks. A robust layerd control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23, 1986.

[14] M. A. Brown. On the logic of ability. *Journal of Philosophical Logic*, 17, 1988.

[15] Paolo Busetta and Kotagiri Ramamohanarao. The BDIM agent toolkit design. Technical report, Department of computer science, Univerisity of Melbourne, Parville, Victoria, Australia, 1997.

[16] Paolo Busetta, Ralph Ronquist, Andrew Hodgson, and Andrew Lucas. JACK intelligent agents - components for intelligent agents in java. Technical report, Agent Oriented Software Pty. Ltd, Melbourne, Australia, 1998.

[17] C. Caleiro, W. A. Carnielli, Marcelo E. Coniglio, Amílcar Sernadas, and Cristina Sernadas. Fibring non-truth-functional logics: Completeness preservation. *Journal of Logic Language and Information*, 12(2):183–211, 2003.

[18] José Carmo and O Pacheco. Deontic and action logics for organized collective agency modeled through institutionalized agents and roles. *Fundamenta Informaticae*, 48:129–163, 2001.

[19] Laurent Catach. Normal multimodal logics. In *In Proc. National Conference on AI(AAAI-88)*, pages 491–495, 1988.

[20] Lawrence Cavedon, Lin Padgham, Anand Rao, and Elizabeth Sonenberg. Revisiting rationality for agents with intentions. In *In proceedings of the Eigth Australian Joint Conference on AI, Canberra*, 1995.

[21] A. Chagorov and M. Zakharyaschev. *Modal Logic*. Clarendon Press, Oxford, 1997.

[22] B. F. Chellas. *Modal Logic, An Introduction*. Cambridge University Press, Cambridge, 1980.

[23] Xiaoping Chen and Guiquan Liu. A logic of intention. In *International Joint Conference on Artificial Intelligence (IJCAI-99)*, 1999.

[24] Paul R. Cohen and Hector J. Levesque. Persistence, intention and commitment. In *In proceedings Timberline workshop on Reasoning about plans and actions*, pages 297–338, 1986.

[25] Paul R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3), 1990.

[26] Marcelo E. Coniglio, Ana Teresa Martins, Amílcar Sernadas, and Cristina Sernadas. Fibring (Para)consistent logics. Technical report, Center for Logic and Computation, Dept. of Mathematics, Instituto Superior Tecnico, 2000.

[27] Marcelo E. Coniglio, Amílcar Sernadas, and Cristina Sernadas. Fibring logics with topos semantics. To appear in Journal of Logic and Computation.

[28] M. J. Cresswell. A henkin completeness theorem for T. *Notre Damne journal of Formal Logic*, 1967.

[29] Daniel C. Dennet. *The Intentional Stance*. The MIT Press, 1989.

[30] Dag Elgesem. *Action theory and modal logic*. PhD thesis, Institute for Philosophy, University of Oslo, 1993.

[31] Dag Elgesem. The modal logic of agency. *Nordic journal of philosophical logic*, 2(2):1–46, 1997.

[32] E. A. Emerson. *Temporal and Modal Logic*. Elsevier, Cambridge, 1990.

[33] E. A. Emerson and J. Srinivasan. *Branching time temporal Logic*. Springer- Verlag, Berlin, 1989.

[34] Ingrand F. F., Michael P. Georgeff, and Anand S. Rao. An architecture for real-time reasoning and system control. *IEEE Expert*, 7(6), 1992.

[35] Kit Fine and G. Schurz. Transfer theorems for multi-modal logics. In *Logic and Reality, Essays in Pure and applied logic. In memeory of Arthur Prior*, pages 169–213. Oxford University Press, 1996.

[36] M. Finger and Dov M. Gabbay. Adding a temporal dimension to a logic. *Journal of Logic, language and Information*, 1:203–233, 1992.

[37] K. Fischer and Muller J. P. & Pischel M. A pragmatic BDI architecture. In *Intelligent Agents II, LNAI-1037*, volume 1037, pages 203–218. Springer-Verlag, 1996.

[38] Michael Fisher. Implementing BDI-like systems by direct execution. In *International Joint Conference on Artificial Intelligence (IJCAI(1)-97)*, 1997.

[39] Abhaya Nayak & Norman Foo, Maurice Pagnucco, and Abdul Sattar. Chnaging conditional belief unconditionally. In *In proceedings of TARK 96 (Theoretical Aspects of Rationality and Knowledge*, pages 119–135, 1996.

[40] Dov M. Gabbay. Fibred semantics and the weaving of logics. part 1. Modal and Intuitionistic Logics. *Journal of Symbolic Logic*, 1996.

[41] Dov M. Gabbay. An overview of fibred semantics and the combinations of logics. In *Frontiers of Combining Systems: Proc of the 1st Int. workshop*, pages 1–56. Kluwer, 1996.

[42] Dov M. Gabbay. *Fibring Logics*. Oxford University Press, Oxford, 1999.

[43] Dov M. Gabbay and Guido Governatori. Dealing with label dependent deontic modalities. In Paul McNamara and Henry Prakken, editors, *Norms, Logics and Information Systems. New Studies in Deontic Logic*. IOS Press, Amsterdam, 1998.

[44] Dov M. Gabbay and Guido Governatori. Fibred modal tableaux. In David Basin, Marcello D'Agostino, Dov Gabbay, Sean Matthews, and Luca Viganó, editors, *Labelled Deduction*, volume 17 of *Applied Logic Series*, pages 163–194. Kluwer, Dordrecht, 2000.

[45] Dov M. Gabbay, A. Kurucz, Frank Wolter, and M. Zakharyaschev. Many-dimensional modal logics: Theory and applications. A preliminary version is available at http://www.dcs.kcl.ac.uk/staff/dg/.

[46] Dov M. Gabbay and Valentin B. Shehtman. Products of modal logics, part 1. *Logic Journal of the IGPL*, 6(1):73–146, 1998.

[47] Dov M. Gabbay and Valentin B. Shehtman. Products of modal logics, part 2: Relativised quantifiers in classical logic. *Logic Journal of the IGPL*, 8(2):165–210, 2000.

[48] P. Gardenfors. *Knowledge in Flux: Modelling the Dynamics of Epistemic States*. The MIT Press, Cambridge, Massachusetts and London, England, 1988.

[49] P. Gardenfors. *Knowledge in Flux: Modelling the Dynamics of Epistemic States*. The MIT Press, Cambridge, Massachusetts and London, England, 1988.

[50] Jonathan Gelati, Guido Governatori, Antonino Rotolo, and Giovanni Sartor. Declarative power, representation, and mandate: A formal anaysis. In Trevor Bench-Capon, Aspassia Daskalopulu, and Winkels Radboudb, editors, *Legal Knowledge and Information Systems*, number 89 in Frontieres in Artificial Intelligence and Applications, pages 41–52. IOS Press, Amsterdam, December, 16-17 2002.

[51] M. Genesereth and N. Nilson. *Logical foundations of Artificial Intelligence*. Morgan Kaufmann, 1987.

[52] Michael P. Georgeff and Anand S. Rao. The semantics of Intention Maintenance for Rational Agents. In *Proceedings of the fourteenth International joint conference on Artificial Intelligence (IJCAI-95)*, pages 704–710, 1995.

[53] Paolo Di Giusto and Guido Governatori. Analytic modal revision for multi-agent systems. In P. Barahona and J.J. Alferes, editors, *Progress in Artificial Intelligence*, volume 1695 of *LNAI*, pages 282–296, Berlin, 1999. Springer-Verlag.

[54] Robert Goldblatt. *Logics of time and computation*. CSLI publications, 1987.

[55] Guido Governatori. Labelled tableaux for multi-modal logics. In P. Baumgartner, R. Hähnle, and J. Posegga, editors, *Theorem Proving with Analytic Tableaux and Related Methods*, volume 918 of *LNAI*, pages 79–94, Berlin, 1995. Springer-Verlag.

[56] Guido Governatori. Labelling ideality and subideality. In Dov M. Gabbay and Hans Jürgen Ohlbach, editors, *Practical Reasoning*, number 1085 in Lecture Notes in Artificial Intelligence, pages 291–304, Berlin, 1996. Springer-Verlag.

[57] Guido Governatori, Jonathan Gelati, Antonino Rotolo, and Giovanni Sartor. Actions, institutions, powers. preliminary notes. In Gabriela Lindemann, Daniel Moldt, Mario Paolucci, and Bin Yu, editors, *International Workshop on Regulated Agent-Based Social Systems: Theories and Applications (RASTA'02)*, number FBI-HH-M-318/02 in Mitteilung, pages 131–147, Hamburg, 12 July 2002. Fachbereich Informatik, Universität Hamburg.

[58] Guido Governatori and Michael Maher. An argumentation-theoretic cahracterisation of defeasible logic. In *Proceedings of the 14th european conference on artificial intelligence (ECAI-2000)*, pages 469–473, 2000.

[59] Guido Governatori, Vineet Padmanabhan, and Abdul Sattar. A Defeasible Logic of Policy-based Intentions. In *AI 2002: Advances in Artificial Intelligence*, LNAI-2557. Springer Verlag, 2002.

[60] Guido Governatori, Vineet Padmanabhan, and Abdul Sattar. A Defeasible Logic of Policy-based Intentions. In *Australian workshop on computational intelligence (AWCL)-02*. Springer Verlag, 2002.

[61] Guido Governatori, Vineet Padmanabhan, and Abdul Sattar. On Fibring Semantics for BDI Logics. In *Logics in Artificial Intelligence: 8th European conference (JELIA-02), Cozenza, Italy*, LNAI-2424. Springer Verlag, 2002.

[62] Guido Governatori and Antonino Rotolo. Brief notes on the axiomatisation of Elgessem's logic of action and ability. Not published.

[63] Afsaneh Haddadi. *Communication and Cooperation in Agent Systems: A Pragmatic Theory*, volume 1056 of *Lecture Notes In Computer Science*. Springer-Verlag, Berlin, Heidelberg, NewYork, 1996. Subseries LNAI.

[64] Joseph Y. Halpern and Yoram Moses. A guide to completeness and complexity for modal logics fo knowledge and belief. *Artificial Intelligence*, 54(3):319–379, April 1992.

[65] D. Harel. Dynamic logic. In *Handbook of Philosophical Logic*. D. Reidel publishing company, 1984.

[66] Edith Hemaspaandra. Complexity transfer for modal logic. In *In Proceedings of the Ninth IEEE symposium on Logic in Computer Science LICS-94*, Paris, France, 1994.

[67] J. Hintikka. *Knowledge and Belief*. Cornell University Press, 1962.

[68] M. Huber. Jam: A BDI-theoretic mobile agent architecture. In *Third International Confernce on Autonomous Agents - Agents 99*, pages 236–243, Seattle, WA, 1999.

[69] G.E. Hughes and M.J. Cresswell. *An Introduction to Modal Logic*. Routledge, London, 1968.

[70] G.E. Hughes and M.J. Cresswell. *A New Introduction to Modal Logic*. Routledge, New York, 1996.

[71] Andrew J. I. Jones and Marek J. Sergot. A formal characterisation of institutionalised power. *Journal of the IGPL*, 4:429–445, 1996.

[72] Hirofumi Katsuno and Alberto O. Mendelzon. On the difference between updating a knowledge base and revisiong it. In Peter Gardenfors, editor, *Belief Revision*, pages 183–203. Cambridge University Press, 1992.

[73] Kurt Konolige. *A Deduction Model of Belief*. Pitman/Morgan Kaufmann, 1986.

[74] Kurt Konolige and Martha E. Pollock. A representationalist theory of intention. In *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-93)*, pages 390–395, 1993.

[75] J. Kozen, D. & Tiuryn. *Logics of Programs*. Elsevier, 1990.

[76] Marcus Kracht. *Tools and Techniques for Modal Logics*. Elsevier, 1999.

[77] Marcus Kracht and Frank Wolter. Properties of independently axiomatizable bimodal logics. *The Journal of Symbolic Logic*, 56(4):1469–1485, 1991.

[78] Saul A. Kripke. Semantical analysis of modal logic. *Zeitschrift fuer Mathematische Logik und Grundlagen der mathematik*, 9:67–96, 1963.

[79] C. Krog. Obligations in multi-agent systems. In *Fifth Scandinavian Conference on Artificial Intelligence (SCAI-95)*, pages 19–30, 1995.

[80] E. J. Lemmon and D. S. Scott. *The Lemmon Notes: An Introduction to Modal Logic*. Blackwell, 1977.

[81] E.J. Lemmon. Algebraic semantics for modal logic II. *Journal of Symbolic Logic*, 31:191–218, June 1966.

[82] C. I. Lewis. *A Survey of Symbolic Logic*. University of California Press, Berkeley, 1918.

[83] C. I. Lewis and C. H. Lanford. *Symbolic Logic*. Appleton Century Crofts, New York, 1932.

[84] Jorge Lobo, Randeep Bhatia, and Shamim Naqvi. A policy description language. In *Proceedings of AAAI-99*. AAAI/MIT Press, 1999.

[85] Alessio Lomuscio. *Information Sharing Among Ideal Agents*. PhD thesis, School of Computer Science, University of Brimingham, 1999.

[86] Franz Baader & Carsten Lutz, Holger Sturn, and Frank Wolter. Fusions of description logics and abstract description systems. *Journal of Artificial Intelligence Research*, 16:1–58, 2002.

[87] Michael J. Maher. Propositional defeasible logic has linear complexity. *Theory and Practice of Logic Programming*, 1(6):691–711, 2001.

[88] Michael J. Maher, Andrew Rock, Grigoris Antoniou, David Billignton, and Timothy Miller. Efficient defeasible reasoning systems. *International Journal of Artificial Intelligence Tools*, 10(4), 2001.

[89] D. C. Makinson. On some completeness theorems in modal logic. *Zeitschrift fur mathemathische Logik und Grundlagen der Mathematik*, 12:379–384, 1966.

[90] Maarten Marx. Interpolation in modal logic. In *Proceedings of Algebraic methodology and software technology (AMAST)*, pages 154–163. Springer, 1999.

[91] Maarten Marx and Carlos Areces. Failure of interpolation in combined modal logics. *Notre Damne Journal of Formal Logic*, 39(2):253–272, 1998.

[92] John McCarthy. Modality, si! modal logic, no! *Studia Logica*, pages 29–32, 1997.

[93] Ben Moszkowski. *Executing temporal logic programs.* Cambridge university press, 1986.

[94] Naoyuki N. Deduction systems for BDI logics using sequent calculus. In *Proceedings of the first international conference on autonomous agents and multi-agent systems (AAMAS-02), Bologna, Italy.* ACM, 2002.

[95] Donald Nute. Defeasible logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3, pages 353–395. Oxford University Press, 1987.

[96] Donald Nute. Defeasible reasoning. In *Proceedings of 20th Hawaii International Conference on System Science*, pages 470–477. IEEE press, 1987.

[97] Lin Padgham and Patrick Lambrix. Agent capabilities: Extending BDI theory. In *Proccedings of AAAI-2000, Austin, Texas, USA*, pages 68–73, 2000.

[98] Vineet Padmanabhan, Guido Governatori, and Abdul Sattar. Actions made explicit in BDI. In *AI 2001: Advances in Artificial Intelligence*, LNAI-2256. Springer Verlag, 2001.

[99] Vineet Padmanabhan, Abdul Sattar, K. Pujari Arun, and Chinmoy Goswamy. Temporal reasoning: A three way analysis. In *Proceedings of the Seventh International Workshop On Temporal Representation And Reasoning (TIME-00)*, pages 183–189. IEEE Computer Society, 2000.

[100] Anand S. Rao. Decision procedures for propositional belief-desire-intention logics. Technical note - 44, Australian Artificial Intelligence Institute, 1993.

[101] Anand S. Rao. Formal models and decision procedures for multi-agent systems. Technical note - 61, Australian Artificial Intelligence Institute, 1995.

[102] Anand S. Rao. Agentspeak(l): BDI agents speak out in a logical computable language. Technical note - 64, Australian Artificial Intelligence Institute, 1996. Also appears in *Agents Breaking Away*, LNCS, Vol. 1038 1996.

[103] Anand S. Rao. Decision procedures for propositional linear-time belief- desire-intention logics. In *Intelligent Agents Volume II, LNAI-1037*, pages 33–48. Springer-Verlag, 1996.

[104] Anand S. Rao and Norman Foo. Minimal change and maximal coherence. In *International Joint Conference on Artificial Intelligence, IJCAI*, pages 966–970, IJCAI-89, 1989.

[105] Anand S. Rao and Michael P. Georgeff. Assymetry thesis and side-effect problems in linear-time and branching-time intention logics. Technical note - 13, Australian Artificial Intelligence Institute, 1991. Also appears in IJCAI-91.

[106] Anand S. Rao and Michael P. Georgeff. Deliberation and its role in the formation of intentions. In *Uncertainity in Artificial Intelligence*, pages 300–307, Seventh Conference on Uncertainity in AI, 1991.

[107] Anand S. Rao and Michael P. Georgeff. Modelling rational agents within a BDI-architecture. In *Principles of Knowledge Representation and Reasoning (KR'91)*. Morgan Kaufmann, 1991.

[108] Anand S. Rao and Michael P. Georgeff. An abstract architecture for rational agents. In *Proceedings of the third International conference on Principles of Knowledge Representation and Reasoning (KR-92)*, pages 439–449, 1992.

[109] M. D. Sadek. A study in the logic of intention. In *Proceedings of the third International conference on Principles of Knowledge Representation and Reasoning (KR-92)*, 1992.

[110] H. Sahlqvist. Completeness and correspondence in the first and second order semantics for modal logic. In *Proceedings of the third Scandinavian logic symposium, Uppsala, Sweden*, 1973.

[111] Klaus Schild. On the relationship between BDI logics and standard logics of concurrency. In *Intelligent Agents - 5, Agent Theories, Architectures and Languages*, pages 47–61. Springer, 1998.

[112] Amiílcar Sernadas, Cristina Sernadas, and Carlos Caleriro. Fibring of logics as a categorial construction. *Journal of Logic and Computation*, 9(2):149–179, 1999.

[113] Amílcar Sernadas, Cristina Sernadas, C. Caleiro, and T. Mossakowski. Categorial fibring of logics with terms and binding operators. In *Frontiers of Combining systems 2*, pages 295–316, 2000.

[114] Amílcar Sernadas, Cristina Sernadas, and M. Ryan. Combining logics: Synchronising and fibring. Technical report, Department of Mathematics, Instituto superior Tecnico, Lisbon, Portugal, 1996.

[115] Amílcar Sernadas, Cristina Sernadas, and A. Zanardo. Fibring modal first-order logics: Completeness preservation. *Logic Journal of the IGPL*, 10(4):413–451, 2002.

[116] Munindar P. Singh. A critical examination of the Cohen-Levesque theory of intentions. In *In Proc. European Conference in Artificial Intelligence, (ECAI-92)*, 1992.

[117] Munindar P. Singh. Semantical considerations on intention dynamics for BDI agents. *Journal of Experimental and Theoretical Artificial Intelligence*, 1998.

[118] G. Sommerhoff. The abstract characteristic of living systems. In *Ystems Thinking*. Penguin, 1969.

[119] Tran Cao Son and Jorge Lobo. Reasoning about policies using logic programa. AAAI-spring symposium on answer set programming, March 26-28 2001.

[120] Toru Sugimoto. A preference-based theory of intention. In Riichiro Mizoguchi and John Slaney, editors, *PRICAI-2000, Topics in Artificial Intelligence*, Lecture notes in AI. Springer-Verlag, 2000.

[121] John Thanagrajah, Lin Padgham, and James Harland. Representation and reasoning for goals in bdi agents. In *Australasian Conference on Computer Science*, 2002.

[122] J. van Benthem. Correspondence theory. In *Hanbook of Philosophical logic, Volume II: Extensions of classical logic*. D. Reidel publishing Co., Dordrecht, 1984.

[123] B. Van Linder. *Modal Logic for Rational Agents*. PhD thesis, Department of Computer Science, Utrecht University, 19th June 1996.

[124] Bruce Vermazen and Merril Hintikka. *Essays on Davidson: Actions and Events*. Clarendon Pres, Oxford, 1985.

[125] Heinrich Wansing. Modality, of course! modal logic, si! *Electronic News Journal on reasoning about Actions and Change*, 2:343–347, 1998. http://www.etaij.org/rac/notes/1988/02/.

[126] D.J. Weerasoorya, Anand S. Rao, and K. Ramamohanarao. Design of a concurrent agent-oriented language. In *Intelligent Agents: Theories, Architectures and Languages (LNAI-890)*. Springer-Verlag, 1994.

[127] Frank Wolter. A counter-example in tense logic. *Noter Damne journal of formal logic*, 37(2):167–173, 1996.

[128] Frank Wolter. Fusions of modal logics revisited. In *Advances in Modal Logic*, volume 1. CSLI Lecture notes 87, 1997.

[129] Frank Wolter. The decision problem for combined (modal) logics. Technical report, Institut fur Informatik, universitat Leipzig, Germany, www.informatik.uni-leipzig.de/ wolter/, September 9, 1999.

[130] M. Wooldridge and N. R. Jennings. *Intelligent Agents- Agent Theories, Architectures, and Languages, LNAI-890*. Springer-Verlag, 1995.

[131] Michael Wooldridge. *The Logical Modelling of Computational Multi-Agent Systems*. PhD thesis, Department of Computation, UMIST, Manchester, 1992.

[132] Michael Wooldridge. Practical reasoning with procedual knowledge (a logic of BDI agents with KNOW-HOW). In *Formal and Applied Practical Reasoning*, 1996.

[133] Roberto Zamparelli. Intentions are plans plus wishes (and more). In *AAAI Spring symposium-93, Technical report*, 1993.

[134] A. Zanardo, Amiílcar Sernadas, and Cristina Sernadas. Fibring: Completeness preservation. *Journal of Symbolic Logic*, 66(1):414–439, 2001.