UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MARLO VIEIRA DOS SANTOS E SOUZA

Choices that make you change your mind: a Dynamic Epistemic Logic approach to the semantics of BDI agent programming languages

Thesis presented in partial fulfillment
of the requirements for the degree of
Doctor of Computer Science

Advisor: Prof. Dr. Álvaro Freitas Moreira
Coadvisor: Profa. Dra. Renata Vieira

Porto Alegre
December 2016

CIP — CATALOGING-IN-PUBLICATION

Souza, Marlo Vieira dos Santos e

Choices that make you change your mind: a Dynamic Epistemic Logic approach to the semantics of BDI agent programming languages / Marlo Vieira dos Santos e Souza. – Porto Alegre: PPGC da UFRGS, 2016.

198 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2016. Advisor: Álvaro Freitas Moreira; Coadvisor: Renata Vieira.

I. Moreira, Álvaro Freitas. II. Vieira, Renata. III. Título.

"No time, make or reason."

— Portishead

"Reason is, and ought only to be the slave of the passions"

— David Hume

ACKNOWLEDGEMENTS

First of all, I would like to thank my parents for their support throughout my studies, for putting up with my moods and hysterics when I was overwhelmed, and celebrating every little accomplishment as their own. You guys are my safe port and my greatest examples.

I would like to thank my advisors Álvaro e Renata for their immense support through my studies. Álvaro, your keen eye and relentless questioning have certainly elevated the quality of my work through these years. Without it, this volume would be a lot less easy to read. Renata, thanks for being always the pragmatical and making me see the forest instead of losing myself in the trees. I would probably still be discussing the meaning of *stability* in Bratman or the ontological/epistemological difference between mental actions and ontic actions if it were not for your advices. Also, my thanks to John-Jules Meyer, which for a whole year was my *de facto* advisor and whose questions have greatly contributed to my work in the last two years. Your advices and criticism have steered me in the right direction numerous times.

To my friends, too many to name here. You guys have been my rock these past 5 years. Whether if it was to hear me complain, to take me out and make me remember there was a life outside my pile of papers to read or to stay for literally hours chatting through phone when I was homesick, you have always been there for me. I thank you for keeping me (almost) sane.

To my colleagues from PUCRS (Clarissa, Aline, Larissa and Denise), from UFRGS (Diego, Fabiane, Jonas, Andrei, Felipe and Marcelo) and from Utrecht (Max, Luora, Hein, Bas and Sjoerd). Thanks for the various discussions - usually accompanied by delicious foods and drinks. You inspire me to be a more rounded person and researcher and you show me that science is better when it is made collectively instead of an individual achievement.

LIST OF ABBREVIATIONS AND ACRONYMS

AAP Abstract Agent Programming Language

ADL Agent Dynamic Logic

AOP Agent-Oriented Programming

BDI Belief-Desire-Intention

CTL Computation Tree Logic

DDL Dynamic Doxastic Logic

DEL Dynamic Epistemic Logic

DPL Dynamic Preference Logic

PDL Propositional Dynamic Logic

PRS Practical Reasoning System

QDT Qualitative Decision Theory

LIST OF FIGURES

CONTENTS

ABSTRACT

As the notions of Agency and Multiagent System became important topics for the Computer Science and Artificial Intelligence communities, Agent Programming has been proposed as a paradigm for the development of computer systems. As such, in the last decade, we have seen the flourishing of the literature on Agent Programming with the proposal of several programming languages, e.g. AgentSpeak (RAO, 1996; BORDINI; HUBNER; WOOLDRIDGE, 2007), Jadex (POKAHR; BRAUBACH; LAMERSDORF, 2005), JACK (HOWDEN et al., 2001), 3APL/2APL (DASTANI; VAN RIEMSDIJK; MEYER, 2005; DASTANI, 2008), GOAL (HINDRIKS et al., 2001), among others.

Agent Programming is a programming paradigm proposed by Shoham (1993) in which the minimal units are agents. An agent is an entity composed of mental attitudes, that describe the its internal state - such as its motivations and decisions - as well as its relation to the external world - its beliefs about the world, its obligations, etc. This programming paradigm stems from the work on Philosophy of Action and Artificial Intelligence concerning the notions of intentional action and formal models of agents' mental states. As such, the meaning (and properties) of notions such as belief, desire, intention, etc. as studied in these disciplines are of central importance to the area. Particularly, we will concentrate in our work on agent programming languages influenced by the so-called BDI paradigm of agency, in which an agent is described by her beliefs, desires, intentions.

While the engineering of such languages has been much discussed, the connections between the theoretical work on Philosophy and Artificial Intelligence and its implementations in programming languages are not so clearly understood yet. This distance between theory and practice has been acknowledged in the literature for agent programming languages and is commonly known as the "semantic gap". Many authors have attempted to tackle this problem for different programming languages, as for the case of AgentSpeak (BORDINI; MOREIRA, 2004), GOAL (HINDRIKS; VAN DER HOEK, 2008), etc. In fact, Rao (1996, p. 44) states that "[t]he holy grail of BDI agent research is to show such a one-to-one correspondence with a reasonably useful and expressive language."

One crucial limitation in the previous attempts to connect agent programming languages and BDI logics, in our opinion, is that the connection is mainly established at the static level, i.e. they show how a given program state can be interpreted as a BDI mental state. It is not clear in these attempts, however, how the execution of the program may be understood as changes in the mental state of the agent. The reason for this, in our opinion, is that the formalisms employed

to construct BDI logics are usually static, i.e. cannot represent actions and change, or can only represent ontic change, not mental change.

The act of revising one's beliefs or adopting a given desire are mental actions (or internal actions) and, as such, different from performing an action over the environment (an ontic or external action). This difference is well recognized in the literature on the semantics of agent programming languages (D'INVERNO et al., 1998; BORDINI; HUBNER; WOOLDRIDGE, 2007; MENEGUZZI; LUCK, 2009), but this difference is lost when translating their semantics into a BDI logic. We believe the main reason for that is a lack of expressibility in the formalisms used to model BDI reasoning.

Dynamic Epistemic Logic, or DEL, is a family of dynamic modal logics to study information change and the dynamics of mental attitudes inspired by the Dutch School on the "dynamic turn" in Logic (VAN BENTHEM, 1996). This formalism stems from various approaches in the study of belief change and differs from previous studies, such as AGM Belief Revision, by shifting from extra-logical characterization of changes in the agents attitudes to their integration within the representation language. In the context of Dynamic Epistemic Logic, the Dynamic Preference Logic of Girard (2008) seems like an ideal candidate, having already been used to study diverse mental attitudes, such as Obligations (VAN BENTHEM; GROSSI; LIU, 2014), Beliefs (GIRARD; ROTT, 2014), Preferences (GIRARD, 2008), etc.

We believe Dynamic Preference Logic to be the ideal semantic framework to construct a formal theory of BDI reasoning which can be used to specify an agent programming language semantics. The reason for that is that inside this logic we can faithfully represent the static state of a agent program, i.e. the agent's mental state, as well as the changes in the state of the agent program by means of the agent's reasoning, i.e. by means of her mental actions.

As such, in this work we go further in closing the semantic gap between agent programs and agency theories and explore not only the static connections between program states and possible worlds models, but also how the program execution of a language based on common operations - such as addition/removal of information in the already mentioned bases - may be understood as semantic transformations in the models, as studied in Dynamic Logics. With this, we provide a set of operations for the implementation of agent programming languages which are semantically safe and we connect an agent program execution with the dynamic properties in the formal theory.

Lastly, by these connections, we provide a framework to study the dynamics of different mental attitudes, such as beliefs, goals and intentions, and how to reproduce the desirable properties proposed in theories of Agency in a programming language semantics.

Keywords: Agent Programming. BDI Agents. Belief Change. Goal Change. Dynamic Epistemic Logic. Formal Semantics.

Dynâmica de atitudes mentais em linguages de programação BDI

RESUMO

Dada a importância de agentes inteligentes e sistemas multiagentes na Ciência da Computação e na Inteligência Artificial, a programação orientada a agentes (AOP, do inglês *Agent-oriented programming*) emergiu como um novo paradigma para a criação de sistemas computacionais complexos. Assim, nas últimas décadas, houve um florescimento da literatura em programação orientada a agentes e, com isso, surgiram diversas linguages de programação seguindo tal paradigma, como AgentSpeak (RAO, 1996; BORDINI; HUBNER; WOOLDRIDGE, 2007), Jadex (POKAHR; BRAUBACH; LAMERSDORF, 2005), 3APL/2APL (DASTANI; VAN RIEMSDIJK; MEYER, 2005; DASTANI, 2008), GOAL (HINDRIKS et al., 2001), entre outras.

Programação orientada a agentes é um paradigma de programação proposto por Shoham (1993) no qual os elementos mínimos de um programa são agentes. Shoham (1993) defende que agentes autônomos e sistemas multiagentes configuram-se como uma forma diferente de se organizar uma solução para um problema computacional, de forma que a construção de um sistema multiagente para a solução de um problema pode ser entendida como um paradgima de programação. Para entender tal paradigma, é necessário entender o conceito de agente. Agente, nesse contexto, é uma entidade computacional descrita por certos atributos - chamados de atitudes mentais - que descrevem o seu estado interno e sua relação com o ambiente externo. Atribuir a interpretação de atitudes mentais a tais atributos é válida, defende Shoham (1993), uma vez que esses atributos se comportem de forma semelhante as atitudes mentais usadas para descrever o comportamento humano e desde que sejam pragmaticamente justificáveis, i.e. úteis à solução do problema.

Entender, portanto, o significado de termos como 'crença', 'desejo', 'intenção', etc., assim como suas propriedades fundamentais, é de fundamental importância para estabelecer linguagens de programação orientadas a agentes. Nesse trabalho, vamos nos preocupar com um tipo específico de linguagens de programação orientadas a agentes, as chamadas linguagens BDI. Linguagens BDI são baseadas na teoria BDI da Filosofia da Ação em que o estado mental de um agente (e suas ações) é descrito por suas crenças, desejos e intenções.

Enquanto a construção de sistemas baseados em agentes e linguagens de programação foram tópicos bastante discutidos na literatura, a conexão entre tais sistemas e linguagens com o trabalho teórico proveniente da Inteligência Artificial e da Filosofia da Ação ainda não está bem estabelecida. Essa distância entre a teoria e a prática da construção de sistemas é bem reconhe-

cida na literatura relevante e comumente chamada de "*gap* semântico" (*gap* em inglês significa lacuna ou abertura e representa a distância entre os modelos teóricos e sua implementação em linguagens e sistemas).

Muitos trabalhos tentaram atacar o problema do *gap* semântico para linguagens de programação específicas, como para as linguagens AgentSpeak (BORDINI; MOREIRA, 2004), GOAL (HINDRIKS; VAN DER HOEK, 2008), etc. De fato, Rao (1996, p. 44) afirma que "O cálice sagrado da pesquisa em agentes BDI é mostrar uma correspondência 1-a-1 com uma linguagem razoavelmente útil e expressiva" (tradução nossa)[1]

Uma limitação crucial, em nossa opinião, das tentativas passadas de estabeler uma conexão entre linguagens de programação orientadas a agentes e lógicas BDI é que elas se baseiam em estabelecer a interpretação de um programa somente no nível estático. De outra forma, dado um estado de um programa, tais trabalhos tentam estabelecer uma interpretação declarativa, i.e. baseada em lógica, do estado do programa respresentando assim o estado mental do agente. Não é claro, entretanto, como a execução do programa pode ser entendida enquanto mudanças no estado mental do agente.

A razão para isso, nós acreditamos, está nos formalismos utilizados para especificar agentes BDI. De fato, as lógicas BDI propostas são, em sua maioria, estáticas ou incapazes de representar ações mentais.

O ato de revisão uma crença, adotar um objetivo ou mudar de opinião são exemplos de ações mentais, i.e. ações que são executadas internarmente ao agente e afetando somente seu estado mental, sendo portanto não observáveis. Tais ações são, em nossa opinião, intrinsecamente diferentes de ações ônticas que consistem de comportamento observável e que possivelmente afeta o ambiente externo ao agente.

Essa diferença é comumente reconhecida no estudo da semântica de linguagens de programação orientadas a agentes (BORDINI; HUBNER; WOOLDRIDGE, 2007; D'INVERNO et al., 1998; MENEGUZZI; LUCK, 2009), entretanto os formalismos disponíveis para se especificar raciocínio BDI, em nosso conhecimento, não provem recursos expressivos para codificar tal diferença. Nós acreditamos que, para atacar o *gap* semântico, precisamos de um ferramental semântico que permita a especificação de ações mentais, assim como ações ônticas.

Lógicas Dinâmicas Epistêmicas (DEL, do inglês *Dynamic Epistemic Logic*) são uma família de lógicas modais dinâmicas largamente utilizadas para estudar os fenômenos de mudança do estado mental de agentes. Os trabalhos em DEL foram fortemente influenciados pela escola holandesa de lógica, com maior proponente Johna Van Benthem, e seu "desvio dinâmico" em

[1] No original, em inglês: *"[t]he holy grail of BDI agent research is to show such a one-to-one correspondence with a reasonably useful and expressive language."*

lógica (*dynamic turn* em inglês) que propõe a utilização de lógicas dinâmicas para compreender ações de mudanças mentais (VAN BENTHEM, 1996).

O formalismo das DEL deriva de diversas vertentes do estudo de mudança epistêmica, como o trabalho em teoria da Revisão de Crenças AGM (ALCHOURRÓN; GÄRDENFORS; MAKINSON, 1985), e Epistemologia Bayesiana (HÁJEK; HARTMANN, 2010). Tais lógicas adotam a abordagem, primeiro proposta por Segerberg (1999), de representar mudanças epistêmicas dentro da mesma linguagem utilizada para representar as noções de crença e conhecimento, diferente da abordagem extra-semântica do Revisão de Crenças *a la* AGM.

No contexto das DEL, uma lógica nos parece particulamente interessante para o estudo de programação orientada a agentes: a Lógica Dinâmica de Preferências (DPL, do inglês *Dynamic Preference Logic*) de Girard (2008). DPL, também conhecida como lógica dinâmica de ordem, é uma lógica dinâmica para o estudo de preferências que possui grande expressibilidade para codificar diversas atiutudes mentais. De fato, tal lógica foi empregada para o estudo de obrigações (VAN BENTHEM; GROSSI; LIU, 2014), crenças (GIRARD; ROTT, 2014), preferências (GIRARD, 2008), etc. Tal lógica possui fortes ligações com raciocínio não-monotônico e com lógicas já propostas para o estudo de atitudes mentais na área de Teoria da Decisão (BOUTILIER, 1994b)

Nós acreditamos que DPL constitui um candidato ideal para ser utilizado como ferramental semântico para se estudar atitudes mentais da teoria BDI por permitir grande flexibilidade para representação de tais atitudes, assim como por permitir a fácil representação de ações mentais como revisão de crenças, adoção de desejos, etc. Mais ainda, pelo trabalho de Liu (2011), sabemos que existem representações sintáticas dos modelos de tal lógica que podem ser utilizados para raciocinar sobre atitudes mentais, sendo assim candidatos naturais para serem utilizados como estruturas de dados para uma implementação semanticamente fundamentada de uma linguagem de programação orientada a agentes.

Assim, nesse trabalho nós avançamos no problema de reduzir o *gap* semântico entre linguagens de programação orientadas a agentes e formalismos lógicos para especificar agentes BDI. Nós exploramos não somente como estabelecer as conexões entre as estruturas estáticas, i.e. estado de um programa e um modelo da lógica, mas também como as ações de raciocínio pelas quais se especifica a semântica formal de uma linguagem de programação orientada a agentes podem ser entendidas dentro da lógica como operadores dinâmicos que representam ações mentais do agente. Com essa conexão, nós provemos também um conjunto de operações que podem ser utilizadas para se implementar uma linguagem de programação orientada a agentes e que preservam a conexão entre os programas dessa linguagem e os modelos que representam o

estado mental de um agente.
Finalmente, com essas conexões, nós desenvolvemos um arcabouço para estudar a dinâmica de atitudes mentais, tais como crenças, desejos e inteções, e como reproduzir essas propriedades na semântica de linguagens de programação.

Palavras-chave: Programação orientada a agents, Agents BDI, Mudança de Crenças, Mudança de Intenções, Lógica Dinâmica Epistêmica, Semântica Formal.

1 INTRODUCTION

Shoham (1993), as McCarthy (1979) before him, has a very pragmatic and externalist vision about what it means to ascribe mental attitudes to a given system - computational or not. To these authors, one may attribute a given mental attitude - say beliefs - to a system as long as its behaviour is coherent with what we identify as these mental attitudes (beliefs) in humans and as long as it is useful to understand the system behaviour. In McCarthy (1979, p. 1) words:

> "To ascribe certain beliefs, knowledge, free will, intentions, consciousness, abilities or wants to a machine or computer program is legitimate when such an ascription expresses the same information about the machine that it expresses about a person. It is useful when the ascription helps us understand the structure of the machine, its past or future behavior, or how to repair or improve it. It is perhaps never logically required even for humans, but expressing reasonably briefly what is actually known about the state of the machine in a particular situation may require mental qualities or qualities isomorphic to them."

In confluence with this view, in discussing theories of agency in multiagent systems research, Wooldridge (2000, p. 14) says that

> If a logic of agency is not computationally grounded, then this must throw doubt on the claim that this logic can be useful for reasoning about computational agent systems. If we really intend our theories to be theories of computational systems, then computational grounding is an issue that must be addressed.

In other words, Wooldridge claims that theories of rationality must be abstractions of computational structures, from the point of view of Artificial Intelligence research.

Agent Programming, or Agent-oriented Programming, is the paradigm for the development of computational systems in which the basic structural units are mental attitudes such as beliefs, abilities, goals, etc. (SHOHAM, 1993). Adopting the views expressed in MacCarthy and Wooldridge's quotations above, to properly describe a programming language as an agent programming language, one needs to provide an interpretation of the language's constructs in terms of a formal theory of Agency. For that, however, first it is necessary to decide on one appropriate theory describing the notions used informally to explain human behaviour.

By definition, an agent is an embodied entity possessing the ability to perceive her environment and act upon it (FRANKLIN; GRAESSER, 1997; WOOLDRIDGE, 2000; SCHLOSSER, 2015)[1]. It is also well-accepted that agents inhabit a dynamic world, in the sense that it changes according to the effects of the actions performed by the agent through her effectors, as well as possibly other external influences on it. As such, an agent must be capable of changing her mental state to adjust her actions to accommodate the perceived state of affairs.

This thesis concerns the change of an agent's mind, or rather how an agent changes

[1]Notice, we don't mean with these two characteristics define the term 'agent', which can vary according to the application, but only provide minimal conditions an entity must satisfy to be called an agent in our discussion

her mind. While this topic can be explored through several different perspectives, we will restrict further our study to focus on computational agents, as studied in the area of Agent Programming. Mainly we will focus on a theory of agency based on Bratman (1999)'s BDI model, in which the actions of an agent can be explained by means of her beliefs, desires and intentions.

The BDI model has influenced several agent programming systems and architectures. One of these architectures, Georgeff and Lansky (1987)'s Practical Reasoning System (PRS) is arguably one of the most influential agent programming architectures in the literature and has been replicated into several different programming languages, such as AgentSpeak (VIEIRA et al., 2007), Goal (HINDRIKS et al., 2001), 2APL (DASTANI, 2008), 3APL (DASTANI; VAN RIEMSDIJK; MEYER, 2005) among others. Despite that, their connection with BDI logics is still not fully understood.

Although many works in the area of BDI reasoning establish the conditions for rational BDI agents to make decisions c.f. (COHEN; LEVESQUE, 1990; RAO; GEORGEFF, 1998; WOOLDRIDGE, 1996; VAN RIEMSDIJK; DASTANI; MEYER, 2009), few of these theories can be formally connected with the programming languages they inspired. One of the reasons for that state of affairs is that few of these theories explore the dynamics of such attitudes in face of a dynamic environment. Much of the work investigating, in a formal setting, how the mental state of an agent changes with performing actions and perception are limited to agent's belief change and mainly treated algorithmically (ALECHINA et al., 2006; MOREIRA; VIEIRA, 2008), or to the relation between intention and commitment (COHEN; LEVESQUE, 1990; SINGH, 1992; RAO; GEORGEFF, 1998).

Interpretations for mental attitudes within the semantics of BDI programming languages have been proposed for specific languages such as AgentSpeak (BORDINI; MOREIRA, 2004), Goal (HINDRIKS; VAN DER HOEK, 2008; DE BOER et al., 2007), etc. We believe, however, that these interpretations are yet far too limited, since they are invariably influenced by implementation decisions or are often limited in expressibility.

An important application of such a theoretical connection, beyond the point of view of semantic clarification, is the ability to specify and check the properties of programs developed in these languages. Several specification and verification logics have been proposed for multi-agent systems (DASTANI et al., 2010; ALECHINA et al., 2011; DASTANI; VAN RIEMSDIJK; MEYER, 2007; VAN RIEMSDIJK; DASTANI; MEYER, 2005; VAN LINDER; VAN DER HOEK; MEYER, 1996). To be of use, however, these logics need to ground their semantics to that of the programming language. This is done by providing translations between an agent

program state and a model of the agent's mental state in these logics.

The connection proposed by theses authors, however, is mainly static since their specification/verification logics are not powerful enough to capture the changes in the program state as changes in the agent's mental state. We believe this limits the expressibility of the language and curbs the study on the dynamic properties of mental attitudes in Agent Programming.

To remedy this limitation, however, we need a framework in which the dynamics of mental attitudes can be represented. The study of changes in the mental state of an agent has, for long, been limited to the study of change in their beliefs. That's because belief change inherits mature solutions from both the Epistemology and Logic.

While change in the agent's beliefs have been extensively studied in Formal Epistemology, see for example the seminal work of Alchourrón, Gärdenfors and Makinson (1985) on Belief Revision, considerably less attention has been devoted to the way other mental attitudes may change and how changes of different mental attitudes may be connected. Some example are the works of Konolige and Pollack (1993), Wobcke (1996), Van der Hoek, Jamroga and Wooldridge (2007), Icard, Pacuit and Shoham (2010) which focus on how to change the agent's plans based on her beliefs and the work of Grant et al. (2010) on postulates for rational mental change. We believe, however, that these treatments of mental attitude change are both poorly connected with the philosophical discussion and are mainly extra-semantic, in the sense that they are mainly established as operations not definable inside the languages used to reason about the agent's mental state. In general, there is no guarantee that this extra-semantic treatment can be computationally grounded, i.e. connected to the computational structures used in Agent Programming, as required by Wooldridge (2000), or how to interpret agent programs using these frameworks.

We propose that using a richer language that can represent both the agent's mental state and mental state changing operations, as has been proposed in recent works in the study of epistemic dynamics (SEGERBERG, 1999; BALTAG; SMETS, 2008), we can provide a better understanding of the mental attitudes encoded in the semantics of agent programming languages.

Epistemic dynamics studies how an epistemic agent, i.e. an entity capable of holding beliefs and knowledge about the world, change her epistemic state as a function of events that affect her understanding of the world, such as announcements, perception, communication, etc. Several formal approaches to epistemic dynamics have been proposed in the literature such as Dynamic Epistemic Logics (DEL) (BALTAG; MOSS; SOLECKI, 1998), Belief Revision Theory (ALCHOURRÓN; GÄRDENFORS; MAKINSON, 1985), Bayesian Epistemology (HÁJEK; HARTMANN, 2010), etc.

We believe that the theoretical framework of epistemic dynamics can be used to study the dynamics of the mental states of the agent in a unified manner. Thus, we can analyse the behaviour of rational agents in a uniform semantics with techniques and postulates well understood in the literature and a mature philosophical support.

Among these formalisms, we believe the approach of Dynamic Epistemic Logics to be the the most exciting for its well-understood semantics, as well as being easily connected with the long tradition on Belief Revision Theory and Bayesian Epistemology. In the context of Dynamic Epistemic Logics, the Dynamic Preference Logic of Girard (2008) seems like an ideal candidate, having already been used to study diverse mental attitudes, such as Obligations (VAN BENTHEM; GROSSI; LIU, 2014), Beliefs (GIRARD; ROTT, 2014), Preferences (GIRARD, 2008), etc.

1.1 Objectives

The main goal of this thesis is to shorten the semantic gap between computational agents and the BDI theory of agency. Particularly, we want to investigate how the semantics of certain agent programming languages can be related to BDI theories not only on a static level, but how the execution of a given agent program can be characterized as dynamic properties of the theory.

In other words, in this work we wish to investigate not only how one particular program state can be understood in a declarative way, but rather how the progression in the execution of an agent program can be characterized by means of transformations on the models representing the agent's mental state. In this way, we can provide a complete picture of how the semantics of a given programming language is related to a theory of agency. This will helps us to explain how the practical reasoning and mental coherency requirements in the BDI theory (BRATMAN, 1999), are/can be encoded in the semantics of a given agent programming language.

To do that, we will use the formal framework of Dynamic Preference Logic (DPL), *a la* Girard (2008), a dynamic modal logic following the Dynamic Epistemic Logic tradition. We choose this formal framework since it has the potential to become a common framework to study different but correlated notions in the areas of Epistemology (BOUTILIER, 1994a; BALTAG; SMETS, 2008), Deontic Logic (VAN BENTHEM; GROSSI; LIU, 2014) and Agency Theory (BOUTILIER, 1994b; LANG; VAN DER TORRE; WEYDERT, 2003). In this logic, we will represent the notions of beliefs, desires and intentions as well as a set of semantic operations encoding common mental actions performed by agents - such as updating her beliefs, or adopting a desire or intention

We will show how to implement agent program states in a agent programming language using an information structure known as priority graph. Exploring the connection proved by Liu (2011) between the preference models used in Girard (2008)'s logic and priority graphs, the relation between an agent program state and the logical model of agent's mental state is automatically achieved. Further yet, we will show how to describe the rules of the language semantics - usually described by means of an operational semantics *a la* Plotkin (2004) - using a set of semantic operations represented in the logic.

The specific goals of this work can be summarised as:

1. To provide a logic that encodes both static and dynamic requirements of mental attitudes as discussed in BDI theory;
2. To provide a set of 'safe' operations on the models, i.e. operations possessing desired representation results, in this logic that can be used to characterize the mental actions involved in practical reasoning;
3. To characterize these operations by means of syntactic operations over the data structures that will be used to implement the agent programming languages;
4. To study the relation between agent specification logics, such as BDI logics, and logic programming semantics by means of the data structures and operations over them.

1.2 Motivation

The formalization of agency and practical reasoning is an ongoing discussion within the Autonomous Agents and Artificial Intelligence communities. Many competing formal frameworks for specification and reasoning about agency have been proposed and implemented as agent programming languages. One of the most influential paradigms in agent specification is the so-called BDI theory (or architecture), which defines an agent as a complex mental entity possessing attitudes such as beliefs, desires, and intentions.

While interpretations for the mental attitudes encoded in some agent programming languages following the BDI architecture have been constructed for some agent-oriented languages, we believe these interpretations to be still very restrictive to properly understand many important properties of the languages and the systems created using them. This belief is founded on the fact that these interpretations are limited to the static level of the language, i.e. they show how a program state can be understood as a declarative model of agency. The problem, however, is that, while we can understand in a declarative way the state of a given program, it is not

possible to specify within the declarative interpretation how the program execution is carried.

We believe that we can solve this limitation by using a dynamic logic of rationality as a framework to interpret agent programs. With a dynamic logic as a semantic foundation of the programming language, we can specify the operations performed in the program execution by means of semantic operations on the logic - giving a complete interpretation of the programming language semantics.

With this study, we aim to better understand the theoretical foundations of Agent Programming. A better connection between agent programming languages and agency theory provides both a better understanding of the semantics of the languages and the systems created with them.

1.3 Structure of this document

This thesis is structured in three parts: the Philosophy, the Logic and the Application. Chapter 2 comprehends the philosophical part of our studies, presenting some of the discussion in the philosophical literature on agency and intentional action. Chapters 3 to 5 constitute the logical part of this thesis. In these chapters, we discuss some of the logical frameworks proposed in the related literature, the logic of dynamic preferences that we will use as a basis to our encoding of mental attitudes and, finally, propose a logic to reason about BDI agent programs. Lastly, Chapter 6 constitutes the application part of this thesis. In this chapter, we show how the formal semantics of BDI agent programming languages can be connected with the semantic framework proposed to reason about agents. More specifically, this thesis is structured as described bellow.

In Chapter 2, we present some of the vast philosophical work on intentions, highlighting the work of Bratman (1999) for its huge impact in the Artificial Intelligence approach to agency. In this chapter, we will present the basic structure of a BDI theory and discuss some topics such as the nature of propositional attitudes, and views such as intentions as plans. This discussion will help us later to understand the notion of intention usually employed in Agent Programming. Further, we will discuss the topic of stability and reconsideration of intentions, specially what Bratman (1999) says on the subject.

Starting the Logic part of the thesis, in Chapter 3, we show how the philosophical intuitions described in Chapter 2 are distilled in the formal frameworks proposed for reasoning about agency. We will focus on three different topics which are strategic in our work: formalization of mental attitudes in Artificial Intelligence and Agent Programming; formal theories

of intention reconsideration and formalization of notions such as commitment and stability for intentions; logics for mental attitudes and their change.

In our study of the formalization of mental attitudes, we focus on presenting some of the most influential proposals for logics of agency and BDI reasoning - specially those focused on Agent Programming. With the work on intention reconsideration, our aim is to point out how an intentional system must behave in the face of dynamic environments. Finally, by studying logics of beliefs and information change, we present the major developments in the field known as dynamic epistemic logics, the semantic framework we will adopt in our study. Dynamic Epistemic Logics (DEL) are a recent and fruitful approach to the study of several related phenomena on Formal Philosophy which have been applied to model problems in Moral Philosophy (VAN BENTHEM; GROSSI; LIU, 2014), Formal Epistemology (BALTAG; SMETS, 2008), Learning Theory (GIERASIMCZUK, 2009) and other areas.

Chapters 2 and 3 comprehend the revision of the literature related to our work. As such, they present a vastitude of works approaching topics related to the meaning, i.e. semantics, of terms such as 'intention' and 'goal.' We believe that this exposition will aid the reader to comprehend the several questions an appropriate theory of agency must consider and how, for the limited case of Agent Programming, these questions can be satisfactorily solved. At the end of each of these chapters, we will present a summary of the problems presented in each chapter and the choices we make in our work.

While Chapters 2 and 3 are related in the sense that they address the same topics by different perspectives, they are independent of each other. We suggest Chapter 2 to be read first. Nonetheless, for the reader more experienced in Logic and Artificial Intelligence, Chapter 3 may actually be of help in understanding the concepts in Chapter 2.

In Chapter 4, we present the Dynamic Preference Logic, as introduced by Girard (2008), a logic in the family of Dynamic Epistemic Logics which will be the formal framework we will use to provide the interpretation of the BDI mental attitudes. We present the representation results of Liu (2011) connecting the possible world semantics of the logic with the implementation-friendly information structures of priority graphs, which, as we will show later in Chapter 6, will play an important role in the grounding the semantics of the abstract agent programming language we propose in this thesis. Further, we provide some results on the representation of some well-known semantic operations on models as transformations in priority graphs. With this, we provide a set of semantic operations that will be included in the logic as dynamic modalities and that can be described by means of priority structures. These operations will later be used to describe the operational semantics of an abstract agent programming

language - tackling the second specific goal of this thesis, as listed in Section 1.1.

After presenting the language and the basic results, in Chapter 5 we specialise the logic of preferences to encode the notions of beliefs, desires and intentions. With this logic, thus, we tackle the first specific goal of our thesis, as listed in Section 1.1. Using the connection between preference models and priority graphs, we present a syntactic model for an agent's mental state, which will be useful to define the operational semantics of an abstract agent programming language in Chapter 6. Further, we present how the operations presented in Chapter 4 can be used to describe changes in the agent's mental state, such as contracting a belief or adopting a goal. With this we tackle the third specific goal listed in Section 1.1.

As we commented, we organize Chapters 3, 4 and 5 in the part of our thesis focused in logical methods in tools. While Chapters 4 and 5 are not directly dependent on Chapter 3, some of the topics discussed in the latter will be relevant to the discussion in the formers. When possible, in Chapter 4 and 5, we will point out these connections to the reader. Chapter 5, however, is directly dependent on what we present in Chapter 4. As such, we advise the reader to follow this order while reading this thesis.

In Chapter 6, we introduce an abstract agent programming language and its semantics by means of a structural operational semantics-like transition system (PLOTKIN, 2004). Later, we show that the semantic structure used to specify an agent state in this language can be understood by means of the priority graphs studied in Chapter 5. As such, if the semantics of constructs of the programming language can be implemented by means of the operations presented in Chapter 5, we are able to immediately provide declarative interpretations of the agent mental state and its change during the program execution. With this, we can finally tackle the last specific goal of our thesis, as listed in Section 1.1.

Chapter 6 corresponds to the part of the thesis regarding agent programming languages. While most of the presentation of Chapter 6 is not dependent on the rest of the thesis, this chapter is the connection between all the parts of the thesis. As such, in Chapter 6, we revisit the discussions of Chapter 2 and also present a connection between the programming language and the logic presented in Chapter 5.

Finally, we conclude our study with some final considerations about our work and the results obtained. We highlight some lateral research developed during the execution of my PhD studies, as well as some related results we achieved in this period. We conclude the work with some pointers for further developments of our proposal, its strengths and limitations.

In Appendix A, we present the full axiomatization of the logic proposed in Chapter 5 together with the encodings for mental attitudes proposed in the chapter.

Part I

The Philosophy

2 INTENTIONS AND THE BDI MODEL

As stated before, this thesis is divided in three parts: a philosophical introduction on intention and practical rationality; a logic-based analysis of intention and, finally, an application of these logical frameworks in the study of agent programming languages. The starting point is the philosophical analysis of the concepts of intention and action. Intentions hold a central position in the Philosophy of Action, specially by the influence of Anscombe (1957). In the study of intentional (or purposive) action, intention as a mental attitude has come to the forefront of the philosophical debate, with many different proposals to explain its relation to intentional action. For this reason, we focus in this chapter in the study of intentions and their relation to intentional action.

In this chapter we discuss different phenomena involved in characterizing mental attitudes as well as in relating intention as mental states and intentional action and the ontological commitments associated with the answers to this problem. It is important to notice that the exact meaning of the terminology presented in this chapter may vary between approaches and, most importantly, some concepts are not uniformly distinguishable throughout the philosophical and logical literature. In both Chapters 2 and 3, we provide our interpretation and formalization of many of the issues discussed here. We will then make an effort to relate our choices with the approaches discussed in these chapters, justifying them. In the summary at the end of this chapter, we synthesize all those choices, clarifying philosophical positions we will adopt in our work.

Through the presentation of the philosophical literature on intention and action, we will be able to identify the main features and properties for a formal theory of intentions. As such, in this chapter we aim to answer three questions that will help us to evaluate the formal theories of intention proposed in the literature and to indicate the properties which will guide our proposal of a theory of intentions for agent programming languages in the next part of this thesis. The questions that will guide our analysis are: What are the properties of intentions? How does an intention relates to other mental attitudes? When does an agent changes her intentions?

Notice that, while we focus on a straightforward approach from the philosophical requirements for the concept of intention to their realization in agent programming languages, we point out that the study of intentions goes on a two-way street. Firstly, it is clear that the studies in artificial intelligence, such as planning, have had a huge influence on the Philosophy of Action - by the work of Bratman (1999). Also, agent programming languages and agency logics provide a toolbox for the philosopher to investigate different phenomena related to in-

tentions. Devices such as plan failure handling (SARDINA; PADGHAM, 2011) and theoretical connections between intention dynamics and belief revision, as proposed by Wobcke (1996), indicate general answers to the problem of intention reconsideration for the philosopher.

In the following, we will first discuss, in Section 2.1, the concept of intention in Philosophy of Action. In Section 2.2, we focus on the BDI paradigm, giving a special attention to the work of Bratman (1999). Finally, in Section 2.3, we will investigate more closely the problem of intentions stability and reconsideration pointing out the Bratman (1999)'s approach and the problems associated with it. In that section, we will present the principles of intention reconsiderations of Mintoff (2004), which we will adopt in our work.

2.1 The concept of intention

The concept of 'intention' is, at the same time, one of the most important and elusive concepts in the Philosophy of Action and Agency theory. It has, in fact, been re-signified several times by different authors trying to give an analytic definition that accounts for its informal meaning in natural language. To understand the reason for the notion of intention to be so difficult to characterize, it is interesting to point out, as Anscombe (1957) does, that the word *intention* has three very different uses in natural language.

The first is that of *intentional action*, or the notion of doing a certain action *A* purposefully as in the sentence "*S is A-ing intentionally*." The second reading is that of *intention with which*, or the purpose behind the execution of a given action, i.e. a state of affairs the agent wish to achieve doing the action, as in the sentence "*S is A-ing so that X*." Finally, the third reading is that of *prospective intention*, or intention for the future, which defines an intention of an agent regarding a state of affairs and not connected with the execution of a specific action, as in the sentence "*S intends X*." Anscombe (1957) defends that a theory of rational action must explain the connections between these 'three sides' of intending, for this enterprise would reveal the true nature of the notion of intention itself.

To make the difference more clear, we will use the simple example of Wilson and Shpall (2012): if a person moves her head, she may do it intentionally, as opposed to an unconscious reflex to hearing a sound close to her. This action can be carried with the purpose of disagreeing with an interlocutor, e.g. shaking her head, or simply shake an insect off her head. Notice that, assuming the person shakes her head to disagree with the interlocutor, her intention to disagree may (and usually does) pre-exist the occurrence of the action. As such, the intention to disagree with the interlocutor is a prospective intention, that was further realized into the action shaking

one's head.

Perhaps one of the most influential works in the recent developments of Philosophy of Action and on the notion of intention is that of Bratman (1999). His work has had a great impact in the area of Artificial Intelligence, and has spanned an area of active research in formalizations and development of computational systems based on agents.

Bratman (1999) proposes that a general concept of intention cannot be properly explained by means of beliefs and desires only, as proposed by other philosophers such as Anscombe (1957) and Davidson (1979). The reason for this is that intentions play an important role as a conduct-controlling attitude and as input to practical reasoning, which cannot be properly explained by means of beliefs and desires. The author proposes that (prospective) intentions are to be thought as *sui generis* conduct-controlling mental attitudes that resist reconsiderations and play characteristic roles as inputs into further practical reasoning to yet further intentions (BRATMAN, 1999, p. 22).

Instead of taking intentional action ("*S is A-ing intentionally*") as a primitive, Bratman (1999) focus on prospective intentions ("*S intends to X*") as mental attitudes. Prospective intentions are those always about future states of affairs one wishes to accomplish, e.g. to graduate, for which no immediate action is necessarily required, but which guide or narrows the possible choices one can make in the future if one wants to achieve it. This last characteristic of prospective intentions, namely guiding one's choices, has a crucial role in Bratman (1999)'s philosophy, as we will see further in the next section.

It is debatable whether there may be forms of pure intending, i.e. prospective intention, in the sense that the object of an intention is a proposition and no action of any kind has been initiated to achieve it. Following Davidson (1979) and others, we will consider prospective intentions as a primitive and not reducible to the notion of intentional action.

Notice that it is not clear how to characterise what an action is. Usually, it is not possible to separate an action, such as raising an arm, from an unconscious bodily movement, such as a reflex that causes one to raise an arm. While the first can be defined as an action, since it is caused and controlled by the agent, it is dubious that the second can also be considered as such.

The Causal Theory of Action is the view that an observable behaviour can only be considered an action (or yet an intentional or purposeful act) if it possesses some psychological cause or it is involved in some psychological causal process to which its execution is the final result (PACHERIE, 2008). In other words, an observable behaviour can only be called action, instead involuntary or reflexive behaviour, when it is the result of some deliberative process based on the agent's psychological state, i.e. her beliefs, her desires, intentions, etc. Some

examples of approaches following this perspective is the work of Davidson (1979), Bratman (1999), Pacherie (2008), among others.

It is necessary here to comment on the notion of mental causation. We are discussing theories involving concepts such as mental attitudes, but the ontological status of such objects is not clear yet. In fact, there are different perspectives in the literature about the ontological status of these elements.

Some approaches, e.g. Davidson (1979), imply some form of materialism, in which mental states and properties are inexistent or just abstractions of actual physical entities and thus mental causation is just physical causation. Others subscribe to a form of modern dualism, where the mental and the physical aspects of an agent exist separately and somehow interact to determine the observable behaviour.

This discussion is of utter importance when investigating human agency. For the case of Agent Programming, the opinions expressed by Shoham (1993) and by McCarthy (1979) both denounce a clear adherence to functionalism (CHURCHLAND; CHURCHLAND, 1981) and, in fact, an implicit form of materialism where mental states and mental attitudes are not but a pragmatic abstraction of physical states. In our opinion, however, the literature on agent programming languages and agent architectures, e.g. (VAN RIEMSDIJK; DASTANI; MEYER, 2009; RAO, 1996; BRATMAN; ISRAEL; POLLACK, 1988; HINDRIKS, 2008; BORDINI; HUBNER; WOOLDRIDGE, 2007), takes an approach more closely related to a form of dualism, in which the mental state of the program is directly referenced and manipulated as language constructs. Ultimately, in computational systems, these two approaches are only different according to the level of analysis one is willing to go through. While logically we can differentiate 'physical' and 'mental' in an agent architecture, and we are encouraged to do so, in the lowest-level all 'mental' activity can be described by means of the physical states of the hardware - by bytes, volts, etc.

Of course, pragmatically, from the point of view of software engineering, this differentiation is useful, specially when agent programs are used to control physical systems, such as robots, in which goal-oriented reasoning and physical control are usually treated separately. Examples of the application of such interactions lie in game development, e.g. the work of van Oijen, Vanhée and Dignum (2012) presents a middleware for agents integrating 'physical' aspects as perception and action and reasoning, or even embodiment in immersive virtual environments (GARAU et al., 2003), where environment perception and reasoning must be integrated to create a more realistic experience.

2.2 The BDI model

The BDI model of rational action is the view in Philosophy of Action that rational action is defined by three distinct and irreducible mental attitudes - namely beliefs, desires and intentions, hence the name BDI model. It has been mainly put forward by the work of Bratman (1999) and the critical developments on his theory, from the Philosophy side, and the work on formalization and implementation of such notions, on the Artificial Intelligence side. The central idea of the BDI model of action is the idea that intention is a *sui generis* mental attitude not reducible to beliefs and desires.

For Bratman (1999), intentions *are* plans that were adopted to achieve a given desire. As the author puts it: "Intentions are, so to speak, the building blocks of such plans; and plans are intentions writ large." (BRATMAN, 1999, p. 8)

Intentions, for Bratman (1999), are constrained by internal and external consistency requirements. Simply put, internal consistency requires that the agent's intentions are consistent with each other, while external consistency, or strong consistency, requires that agent's intentions are consistent with her beliefs.

Additionally, the author requires that intentions conform to principles of means-end coherence, i.e. that the agent believes the plans she adopted to achieve a desire X are effective means to achieve it.

On the relationship between intentions and other mental attitudes, it is standard view that intending something entails *desiring*[1] it. Some, however, argue that intention may arise from obligations and not intrinsic desire. In response to such a criticism, one can simply argue that intention always entails desiring that is "either intrinsic, extrinsic or partially both" (AUDI, 1973).

Some, as Davidson (1979), argue further that intending implies overwhelming desire, in the sense that an intention is an all-out unconditional judgement that an action is desirable. This amounts to the requirement that intending to A implies that A is more desirable than any alternative. Bratman (1999) proposes an interesting critique for this view of intention, showing that in such a case, choosing between two equally desirable alternatives would be impossible and nonetheless we do it everyday. Contrary to Davidson, Bratman (1999) requires only that an

[1] Notice that we are using desire as a general motivational attitude here. It is common in the literature to describe this requirement as 'inteding something entails *wanting* it', as *wanting* is not as conceptually charged as *desiring*. Since we are not concerned with any functional differentiation of motivational attitudes, we will only use desires in our work. Notice that different motivational attitudes, such as desires, obligations, etc. may have different roles in reasoning and conflicts between these attitudes have been studied in the context of BDI reasoning by Broersen et al. (2001, 2002).

intention is only not less desirable then its alternatives.

The relationship between intending and believing is a more controversial one. It is a common requirement for one's intentions to be consistent with one's beliefs, in the sense that if one intends to *A*, i.e. to perform a certain action *A*, she must believe she can *A*, or that it is not impossible to *A* (AUDI, 1973; DAVIDSON, 1979; BRATMAN, 1999). In fact, this is the position adopted by Bratman.

This requirement, however, seems to be inconsistent with a common position held about prospective intentions and intentional action, known in the literature as the Simple View. The Simple View of intention states that to do a certain action intentionally, an agent must intend to perform such action, i.e. to have a prospective intention to perform that action.

As an example of the incompatibility of the Simple View with the strong consistency requirement, Bratman (1999) proposed the thought experiment known as the Video Game Puzzle. The Video Game Puzzle is a thought experiment in which a player is presented with two similar video games that she will play simultaneously, one with each hand. In the video game, the player is requested to guide a missile into a target. It is assumed that the task is not a trivial one to achieve and the person is ambidextrous and skilled at the game. The video games are linked so that the person can achieve the goal at either individual machine, but not at both at the same time and the person has knowledge about this restriction.

It is not difficult to argue that if the player manage to hit a missile on either target, she has done so intentionally. Also, it is arguably consistent to try to achieve both at the same time, since the agent has low confidence that she will do it at either one - given that the task is not a trivial one. Nonetheless, by requirement of consistency between intentions and beliefs, the agent cannot simultaneously intend to hit the target at both video games.

The author proposes the solution to this problem by negating the Simple View and stating the agent does not intends to hit both targets but only to try to hit them, in the sense that the agent endeavour to hit them, or posses a guiding desire to hit them.

Bratman's solution to the video game problem is often criticized, firstly because it is not clear the difference between intending to *A* and intending to try to *A*. More importantly, as McCann (1991) defends, this approach implies in the creation of several intention-like attitudes in the agent's mind. Another important criticism to Bratman's solution is that it does not provide a way for relating intentional action and prospective intentions, as required by Anscombe (1957).

As Agent Programming has been influenced by Bratman's work, we will see that these philosophical problems may sometimes be reproduced in programming languages and systems proposed in the area. To deal with these problems, several weakenings of Bratman's theory

have been proposed. Some, as Cohen and Levesque (1990), maintain the strong consistency between intentions and beliefs and solve the puzzle by disallowing an agent to perform the actions of trying to achieve both targets at the same time. Dastani et al. (2003), on the other hand, weaken Bratman's consistency requirements, while maintaining the Simple View. In each case, however, the Simple View is commonly maintained in order to provide a coherent understanding of intending and avoid multiple intention-like attitudes.

Perhaps one of greatest contributions of Bratman's account on intentional action is to point out the function of intentions in reducing the search space of practical reasoning for resource-bound agents, i.e. in reducing the possible courses of action that the agent considers in order to achieve desire. Bratman (1999) requires that intentions are conduct-controlling. In his theory this means two things. First, his intentions as plans theory is a causal theory of action, and thus intentions are effective causes of action. More importantly, perhaps, is the fact that, given the strong consistency requirements, currently held intentions constrain the space of entertainability for new intentions, or as Bratman (1999) says it, intentions act as a "window of admissibility" for new intentions. The latter is an important characteristic since, as Bratman, Israel and Pollack (1988) emphasize it, it is a pragmatical restriction that allows resource-bounded rational agents to function without the overwhelming need to reconsider her decisions at each step.

The topic of intention stability and reconsideration is a much more delicate problem. Bratman (1999) views on intention reconsideration are relatively vague. While the author admits that fanatical adherence to intentions cannot be considered rational, he has very few explanations on how to conciliate the necessary stability of intentions in order for them to be effectively a conduct-controlling attitude and the ability of the agent to revoke her previous decisions. This problem will be exclusively discussed in in Section 2.3.

2.3 Changing intentions: stability and reconsideration

A central point in Bratman (1999)'s theory is the property of *stability* or *inertia* of intentions. The author uses this property to delineate the differences between the roles of intention and desires in the causal process leading to actions. As already pointed out, Bratman (1999) requires that intentions are conduct-controllers, meaning that they are effective causes for the future action. This is achieved by an intention not by directly causing one to act in the future - in the sense of a spooky action at a distance from physics - but by (usually) preserving prospective intentions until action execution. As such, this preservation of intentions until the time to act is

what Bratman (1999) calls the stability of an intention.

To Bratman, desires only potentially influence behaviour, as in the example given by the author: one may desire to have a milkshake for lunch and not fall into irrational behaviour by not having one, since this desire may have to be weighted against others, like the desire of losing weight. The role performed by intentions, however, goes beyond the simple influence of behaviour. Intentions control one's conduct. Once I intend to have a milkshake for lunch, and in the case of not changing my intentions until lunchtime, I will not re-consider the pros and cons of having it.

This characteristic of intentions is explained by Bratman by the notion of *commitment*. Intentions, as conduct-controllers, involve commitment to the action intended, meaning that come the time to act, if my intention has been maintained, I shall perform such action.

This particular aspect of intentions - namely the related notions of *commitment* and *stability* - have received a great deal of attention in the formal work stemming from Bratman's philosophy. In fact, in the work of Cohen and Levesque (1990), the authors argue that the notion of intention can be subsumed by *choice* and *commitment*. These authors impose such strong requirements in the commitments of an agent to her intentions that, once held, an intention may only be dropped if the agent has achieved it or no longer believes it to be achievable - what was dubbed single-minded commitment in the literature.

It is not the case that intentions in Bratman's theory are irrevocable or even as stable as proposed by Cohen and Levesque (1990). Quite the contrary, in fact, as can be viewed in the following statement:

> It is not just that prior intentions resist reconsideration in the way diamonds resist being scratched. Rather, along with this tendency of prior intentions go associated norms of practical rationality - norms that concern the rationality of reconsideration and nonreconsideration of prior intentions. (BRATMAN, 1999, p. 60)

Apart from the case upon which the change in one's mind violates the strong consistency criteria established by his theory, i.e. either the agent cease to desire the expected result of her plans or comes to believe she is unable to achieve them, his work can be considerably vague about what are the normative forces that guide reconsideration.

In fact, Bratman (1999) suggests the existence of different forms of reconsideration. First of them is the nonreflective reconsideration, which the author argues to be carried not by means of any direct reflection on the matter of reconsidering an intention, but rather by means of 'certain underlying habits, skills or dispositions' (BRATMAN, 1999, p. 60). The second form of reconsideration, as the author puts it, is a form of policy-based reconsideration in which, while not deliberating on the desire-beliefs reasons for reconsideration, one explicitly reflects

whether to reconsider by means of applying some general policy about when to reconsider. The third case is the second-order deliberation of whether to reconsider given the associated (physical, emotional, etc.) costs regarding the reconsideration process.

Bratman argues that reconsiderations of the latter case are rare, considering the limits of resource-bounded agents like us and the associated costs with such second-order deliberation. The process of this kind of deliberation involves retracting a given intention and weighting its belief-desire reasons.

It is clear from this that Bratman favours a form of rule utilitarism in his theory of intention stability, that is a form of reasoning in which the agent has two sources of reasons to decide - namely the belief-desire reasons and the reconsideration policies. As such, Bratman's approach is subjected to the well known criticism from Smart (1956) to the irrationality of this approach. Smart's criticism consists of pointing out that, in some cases, these reasons are inconsistent, in the sense that the utilitarian justification for the non-reconsideration of a given intention may conflict with the utilitarian justification for its reconsideration given a more attractive alternative. To better illustrate this case, consider the following example - adapted from one presented by Smart (1956).

A time traveller goes back in time to Germany in the year of 1938. Walking by a river close to Berchstegaden, he sees a black haired man with the signature moustache of Adolf Hitler, who is indeed drowning in the river. Not knowing who it is, unquestionably - by application of the moral rule to save any drowning man - it is undoubtedly rational to save Hitler. The question that lies is: being well aware of who it is that is drowning, is it rational for the time traveller to save Hitler?

Rule utilitarism would argue that yes, it is rational, given that the rule to save any drowning man produces, in most cases, good results. From a pure utilitarian ground, however, knowing the consequences of saving Hitler, it would be an indefensible action[2]. Smart (1956) argues that to stick to the moral rule, despite knowing the consequence of one's actions, is an act of rule worship.

In response to this problem, Bratman (1992) argues that a policy for reconsideration has to "as much as possible, issue in reconsideration in all and only would-change/worth-it cases[3]", and particularly "does not apply to cases in which it is obvious to the agent that this is a would-change/worth-it case" (BRATMAN, 1992, p. 11). A further requirement from Bratman

[2]Not considering, of course, the consequences of a temporal paradox. As a whovian, however, I have perfect faith in the idea that time can be changed (sometimes).

[3]For Bratman (1992), these are the cases in which if an agent were to reconsider, she would change her previous decision and it would be worth it, by utilitarian reasons.

(1992) is that such a reconsideration habit/policy must "explicitly include [...] in this basis the impact of such habits of reconsideration on an agent's ability to benefit from forms of [social] coordination."(BRATMAN, 1992, p. 11)

It is not clear, however, how one is supposed to construct such a policy, especially since to know what are the cases one should reconsider, except in some obvious cases, one has to reason on whether one's intentions are the most beneficial - i.e., one must reconsider their intentions. In other words, to be completely sure that the current intentions are the most beneficial - in utilitarian terms - one needs to consider all the utilitarian reasons for these intentions and compare them with the alternatives, which amounts to the process of deliberation based on utilitarian grounds.

For the project of specification of these policies in Artificial Intelligence, however, Bratman (1992, p. 11) defends the adoption of full rule utilitarism and, consequentially rule worship, stating that "general strategies of reconsideration would be close enough to optimal." In fact, in the proposal of the IRMA architecture, the authors claim that:

> [...]one of the jobs of the robot designer is to construct the filter override mechanism so that, other things equal, it minimizes the frequency in which the agent will be [in situations such that she reconsiders when it is not worth it or don't reconsiders when she should have].(BRATMAN; ISRAEL; POLLACK, 1988, p. 19)

Defending the unrestricted rule utilitarist approach, Mintoff (2004) points out that:

> How much cognitive processing (consideration and monitoring) should one perform regarding A? To determine this, we need to balance the costs of processing against the risk of making mistakes. While the issue of whether to A is open, only certain "important" factors (needing specification) relevant to whether A maximizes utility are considered, and it is only at this time that substantial processing costs are incurred regarding A. It is at this time that factors relevant to whether A maximizes utility are relevant to a rational agent's intention to A. The agent then decides to A (or not), thereby closing the issue and ceasing to incur these costs. Thereafter the issue is closed, and only certain "important" changes (needing specification) in the circumstances of A are monitored for relevance to whether A continues to maximize utility, and only the occurrence of such changes prompts reopening the issue, and reconsideration about A. (MINTOFF, 2004, p. 404)

To this author, rule utilitarism can be defended for cases as the time traveller knowingly saving Hitler by the fact that the decision to follow the moral rule is based on utilitarian grounds, and as such believed to be utility maximising. Mintoff (2004), thus, denies that the result of rule utilitarian reasoning and consequentialist utilitarian reasoning can diverge - for the rule is taken as a reason influencing the decision itself. The author argues that the reconsideration, thus, is only justifiable in the cases in which an "important change" occurs in the agent's state of mind. This is the most interesting part of his approach, in our opinion, since the author then delineates what are the nature of such changes.

To explain what kind of changes may incur in intention reconsideration, firstly Mintoff

(2004) introduces the following a maintenance principle for intentions:

> (B) If (i) you rationally believe at t_0: that p_1 and p_2 are so; that p_1 is a reason to judge that q at t_2, and p_2 a reason not to judge this; and that p_1 is a stronger reason than p_2, (ii) at least partly on the basis of these beliefs, you rationally judge at t_0 that q at t_2, (iii) up to $t_1(\leq t_2)$, you continuously and rationally hold these belief about p_1 and p_2, and (iv) up to $t_1(\leq t_2)$, you continuously hold the belief that q at t_2, then your belief at t_1 that p_2 is insufficient reason for you to reconsider your belief at t_1 that q at t_2. (MINTOFF, 2004, p. 413)

The principle (B) above states that if, at the time of intention selection, the reasons the agent held for not pursuing such an intention are not strong enough to prevent the pursuit of it, then they are not enough reason for an agent to reconsider the intention at a later time. As a positive rule for intention reconsideration, this principle delineates that for an agent to reconsider an intention, either the agent has to come to believe in a strong reason not to pursue it or one of the supporting beliefs of the intention, i.e. one of the reasons supporting its adoption, to be revoked.

The second principle presented by Mintoff (2004) is an intention and decision counterpart to principle (B) above.

> (I) If (i) you rationally believe at t_0: that p_1 and p_2 are so; that p_1 is a reason to decide to A at t_2, and p_2 a reason not to decide this; and that p_1 is a stronger reason than p_2, (ii) at least partly on the basis of these beliefs, you rationally decide at t_0 to A at t_2, (iii) up to $t_1(\leq t_2)$, you continuously and rationally hold these belief about p_1 and p_2, and (iv) up to $t_1(\leq t_2)$, you continuously hold the intention to A at t_2, then your belief at t_1 that p_2 is insufficient reason for you to reconsider your intention at t_1 to A at t_2. (MINTOFF, 2004, p. 417)

The same considerations sketched for principle (B) can be applied to principle (I) *mutatis mutandi*. As such, principle (I) states that to reconsider an intention to A, it suffices for an agent to either come to change her heart about her desire supporting the intention to A or be swayed by a strong enough desire for an alternative, i.e. a desire supporting not to A.

It doesn't seem plausible to us, as Mintoff (2004) proposes, that these reconsideration policies can count as utilitarian reasons for taking an action, in Mintoff's defence of rule-utilitarianism. While a decision rule may be utility maximizing, not all instances of that rule result in a utility maximizing consequence. As such, we do not accept the assumption that reconsideration policies are reasons themselves involved in the selection of an intention, but we believe the principles (B) and (I) for intention reconsideration are an interesting proposal to define what Bratman (1999) calls *prima facie* triggers for reconsideration. In fact, we believe these requirements provide a coherent explanation of what constitutes a reason for reconsideration based on utilitarian grounds.

As Mintoff (2004) points out, since all known possibilities have already been considered in the process of intention formation, i.e. in selecting a utility-maximizing desire, these alterna-

tives may not become reasons for reconsideration, unless of a change in the mental state of the agent providing new evidences to consider about some alternative.

Mintoff (2004)'s view on intention reconsideration is consistent with Castelfranchi and Paglieri (2007)'s constructive requirements for the reconsideration of an intention. In the latter approach, however, the authors explicitly specify what kind of beliefs and desires count as reasons for an intention - a subject Mintoff (2004) does not addresses in his account.

Castelfranchi and Paglieri (2007) propose a Belief-Desire theory of intentions and intention reconsideration, which they call a constructive theory of intentions. By a Belief-Desire theory of intentions, we mean that their approach does not subscribe to the BDI model's view that intentions are a *sui generis* mental state, but a complex mental attitude composed of desires and beliefs. In fact, in Castelfranchi and Paglieri (2007)'s theory, intentions are desires[4] supported by a web of beliefs in a very specific manner.

The most prominent characteristic of Castelfranchi and Paglieri's account for intentional action is, perhaps, the strong relationships between desires and beliefs. The authors postulate that beliefs both support and justify desires. In the process of intention formation and execution, it is only trough the presence or absence of pertinent beliefs that a desire may come to be developed into a plan to be carried out by the agent. As such, in this framework an intention may only be reconsidered if either one of its supporting supporting beliefs is dropped of if either beliefs regarding (better) alternate courses of action are adopted.

To specify the process through which a desire is developed into an intention, the authors propose several different types of beliefs, according to their functional role as justifications for desires. For example, an agent may hold a belief that a desire to eat ice cream is preferable to a desire to eat a piece deep-fried tofu, which the authors classify as a preference belief, or a belief that to eat ice cream, the agent must go buy some ice cream, which they classify as a means-end belief. These kinds of beliefs are organized into a cognitive mental process, called goal life-cycle by the authors, which specifies when a given belief is relevant to adopt or reject a given desire.

In our opinion, although architecturally coherent and theoretically illuminating, Paglieri and Castelfranchi's approach has a philosophical problem: motivational attitudes are meaningless, other than for their functional requirement of representing a state of affairs. The de-differentiation of these attitudes into the simple concept of 'goal' proposed by the authors added

[4]The authors employ the terminology of goal, criticizing the term 'desire' as intrinsically endogenous and hedonistic attitudes, while the notion of goal is more general encompassing not only desires but obligations, etc. We will use desires in this description, since our notion of desire is a general motivational attitude, comprising both endogenous attitudes such as wants and exogenous attitudes such as obligations.

to the heavy reliance on beliefs deflates the very meaning of the motivational attitudes involved in the architecture. Also, we believe this approach to also obscure the very notion of belief. Since, in their framework, beliefs are taken to represent both the motivational - through preferences, utilities and instrumentality - and informational aspects of practical reasoning, it is not clear the form nor the content of the agent's beliefs[5].

Notice nonetheless that, while motivated by very different philosophical standpoints in regards to intentions, Castelfranchi and Paglieri (2007)'s position regarding the principles guiding intention reconsideration are very similar to that of Mintoff (2004). We believe these to be promising basic principles for intention reconsideration since they clearly delineate what has to change in the agent's state of mind to cause her to reconsider her intentions. For that, we will adopt principles (B) and (I) above a the basis to evaluate the dynamics of intentions in the framework proposed in Chapter 5 that we will use in our work.

2.4 Summary of the chapter

In this chapter, we discussed the main concept that will be used in our work, namely that of intention. First, in Section 2.1, we present the very concept of intention in Philosophy of Action and some problems related to its characterization and its value in describing rational action. In Section 2.2, we present the BDI paradigm focusing on the work of Bratman (1999), the philosophical framework we adopt ion this work. Finally, in Section 2.3, we discuss some philosophical discussion regarding the concept of intention stability and the requirements for intention reconsideration - which will be of great importance for our approach on the codification of mental attitudes and their connection with agent programming languages. In this section, we give particular attention to the reconsideration principles of Mintoff (2004), which we will adopt in our work.

As stated before, the aim of this chapter was to identify the main features and properties for a formal theory of intentions. For that, we proposed three questions to guide our exploration of the philosophical literature: What are the properties of intentions? How does an intention relate to other mental attitudes? When does an agent changes her intentions?

[5]Particularly, it is not clear to us how to reconcile both the propositional and utilitarian nature of the different beliefs into a coherent representation for beliefs. While Castelfranchi (2013) discusses the 'quantitative' and 'qualitative' dimensions of mental attitudes, the problem he addresses is different, in the sense that the 'qualitative' dimension in his work represents the 'propositional content' of the mental attitude and the 'quantitative' dimension represents the degree of certainty/urgency associated with it. Here, however, the content itself of a belief may be quantitative - i.e. utilitarian - such as in the case of beliefs about cost and beliefs about preferences, or it may be propositional, such as for beliefs about what the agent can do and beliefs about preconditions.

We now must take some time to enumerate the positions we will adopt in our work. In our study of BDI mental attitudes and Agent Programming in Chapters 5 and 6, we adopt as a philosophical background Bratman's intentions as plans paradigm, adopting his consistency requirements, i.e. the requirements that (i) intentions are consistent with each other and (ii) consistent with her beliefs, (iii) intentions imply desiring and (iv) intentions imply some belief in their achievability, i.e. the agent believes she will at least try to achieve her intentions.

Although there are several critics to Bratman's position - and some valid criticism to his requirements in particular (c.f. the discussion about the video game puzzle in Section 2.2) -, the BDI model is both hugely influential in Agent Programming and a holds a great computational appeal, due to its intrinsic relation with traditional computational problems such as planning and its focus on resource bounded rationality. At one side, this relation to traditional computational problems allows us to apply a well-studied framework, such as planning, to implement Agent Programming, while the focus on resource bounded rationality allows us to curb the complexity of some of these problems.

Regarding intention reconsideration, we will adopt Mintoff (2004) principles, i.e. that the agent has reasons to reconsider an intention if, and only if, there is a change in her mental state affecting the beliefs and desires supporting it. While we disagree with Mintoff's defense of rule utilitarianism that provides the foundation to his principles, we believe that the principles itself hold true.

First, since for adopting an intention the agent must consider the utilitarian reasons motivating its adoption, i.e. the beliefs and goals supporting the intention, these attitudes that were already considered in the adoption process may not become a reason to reconsideration - otherwise they would prevent the adoption of the intention in the first place. More yet, since intentions limit the space of entertainability for new intentions, i.e. since once holding an intention the agent actually stops considering alternative courses of action which are incompatible with it, the reconsideration must be triggered not by the existence of alternative courses of action for the agent, but by a significant change in her dispositions towards these alternatives. This is a position also defended by Castelfranchi and Paglieri (2007). As such, only changes in the agents state of mind concerning a certain intention are valid reasons to trigger its reconsideration in a principled manner.

We believe these are comprehensive properties and principles to guide us through the study of formal theories of intention in Chapter 3 and through the development of a formal theory which will be the foundation for our work, in Chapters 5 and 6.

Part II

The Logic

3 STATE OF THE ART IN FORMAL THEORIES OF AGENCY

In this chapter, we will present the state of the art in three topics that deeply affect our work: the semantics of mental attitudes, as understood by perspective of the Agent Programming perspective with a special focus on intentions; the work on mental attitude dynamics, specially intention revision; and the work on belief and correlated information change, particularly the work on Dynamic Epistemic Logic concerning the codification of some mental attitudes, such as beliefs and obligations. The recent developments in these three areas will guide our proposal for a logic of rationality, presented in Chapter 5, that will be used to provide a declarative semantics for an abstract agent programming language.

As for Chapter 2, in this chapter we present several concepts related to the formalization of mental attitudes. We point out that the meaning of some of these concepts varies according to the approach and the application intended by the formalization. As it is true of any formalization of philosophical concepts, many distinctions are lost in the transition between the philosophical and the logical languages. In our presentation, we will attempt to clarify the conceptual similarities and differences of the approaches overviewed in this chapter to the best of our abilities. To do so, we may sacrifice the original terminology adopted by the authors, when necessary, in favour of a clear conceptual separation between the terms used here. As before, we will make an effort to relate our choices with the approaches discussed in this chapter, as well as in Chapter 2.

In a sense, in this chapter, we study how the philosophical discussions presented in Chapter 2 have been reproduced in formal theories of reasoning and in programming languages for Agent Programming. Section 3.1 discusses formal representation of mental attitudes with a focus on intentions, encompassing the discussions presented in Sections 2.1 and 2.2 of the previous chapter; Section 3.2 discusses formal models for intention reconsideration, akin to the discussion of intention stability and reconsideration in Section 2.3; and finally in Section 3.3, we make a brief presentation of the basic notions of Dynamic Epistemic Logic, a formal framework recently employed to study the dynamics of several different mental attitudes and their connection, such as beliefs (BALTAG; SMETS, 2008), obligations (VAN BENTHEM; GROSSI; LIU, 2014), preferences (GIRARD, 2008), etc.

3.1 Semantics of mental attitudes and Agent Programming

The works on formalizations of notions such as belief, desire or goal are too numerous for us to be able to expose and analyse all of them. In this section, we will concentrate on some of those works which are focused or closely related with Agent Programming. Particularly, we will focus on some of the most influential work on the codification of pro-attitudes, such as desires, goals and intentions, using propositional modal logics.

The seminal work of Cohen and Levesque (1990) was the first work, to the best of our knowledge, that tried to formalize the notion of intention as Bratman (1999) defines it. In their work, the authors present a desiderata for a logic of intentions and describe a propositional multimodal logic which they use to define intentions. Cohen and Levesque's desiderata for a theory of intention can be summarised in the following statements (COHEN; LEVESQUE, 1990):

1. Intentions normally pose problems for the agent; the agent needs to determine a way to achieve them;
2. Intentions provide a "screen of admissibility" for adopting other intentions;
3. Agents "track" the success of their attempts to achieve their intentions;
4. The agent believes her intentions are possible;
5. The agent does not believe she will not bring about her intentions;
6. Under certain conditions, the agent believes she will bring about her intention;
7. Agents need not intend all the expected side-effects of their intentions.

The notion of intention in Cohen and Levesque (1990) has been thoroughly analysed in the literature, most notably by Singh (1992). Singh (1992) claims that Cohen and Levesque (1990)'s encoding of intentions does not match up to the supporting intuitions of those authors and leads to counterintuitive behaviour by the agents.

Rao and Georgeff (1998) describe the logic BDI-CTL an extension of the temporal logic CTL (EMERSON; HALPERN, 1985) by introducing modalities for belief, desires and intentions. The BDI-CTL has been an major development as a semantic framework for BDI reasoning.

In the logic BDI-CTL, the modalities for belief, desire and intention are normal modalities defined over a possible world model. The main difference to the approach of Cohen and Levesque (1990) is that a world, in their model, is a branching time structure on which the argument of the intention - a temporal formula - is evaluated. This allows the representation of

several agent behaviours, e.g. *realist agents* that only desire states of affairs they believe to be possible, instead of the constitutive approach taken by Cohen and Levesque (1990), in which the relationship between the mental attitudes is an integral part of their definition of intention itself.

Although of undeniable theoretical importance, Rao and Georgeff (1998)'s work is often criticized for the difficulty to connect these temporal BDI models proposed with the actual implementation of the notions of belief, desire and intention in computational systems. This difficulty lies in providing a way to translate between these models and computational constructs such as plans and sets of goals and facts, which are commonly used to describe the agent as a computer program.

From the perspective of Decision Theory and Planning, we have the work of Boutilier (1994b) and of Thomason (2000). These approaches apply non-monotonic reasoning to encode desires, with the aim to provide a declarative interpretation for them as used in planning. Boutilier (1994b)'s work stems from his study in non-monotonic reasoning for Belief Revision. The author presents an interpretation based on Kripke models for beliefs and desires, instead of the usual utilitarian framework of Decision Theory. Thomason (2000), on the other hand, presents an interpretation for beliefs and desires for the area of planning based on the notion of default (REITER, 1980).

These two proposals are interesting for our study of the semantics of mental attitudes for Agent Programming since it is widely recognized that human reasoning is not just deductive, but also defeasible (POLLOCK, 1987). As such, a defeasible understanding of mental attitudes may provide a more natural interpretation for the phenomena associated with these notions. In our formal framework, presented in Chapters 4 and 5, we will incorporate these concerns by providing a conditional logic for these attitudes, i.e. a non-monotonic logic based in preference models.

As noted by Dastani, Hulstijn and Van der Torre (2001), the notion of desire in Agent Programming differs from other areas such as Qualitative Decision Theory (QDT). In QDT, desires commonly have a double nature: they embed both a motivational nature, i.e. what is preferable or desirable, but also a deliberative nature, i.e. what the agent *chooses* to do. Desires in Agent Programming, however, are usually treated as possible alternatives the agent may pursue.

In fact, it is Bratman (1999) in his philosophy of action that separates these two aspects (motivational and deliberative roles) as different mental attitudes, namely desires and intentions. Intentions for Bratman (1999) have the role of preventing continuous reconsideration of

competing alternatives, allowing thus a resource bounded rational agent to carry on her actions. It is, we believe, in the requirement of *resource bounded rationality* that Bratman's theory and QDT differ: while QDT presupposes an ideal reasoner - heritage from its decision-theoretic roots-, Bratman (1999)'s theory considers *resource bounded* agents such as humans.

In the area Agent Programming, there is a common terminological confusion about the concepts of desire and goal. Both terms have a historical overload in Artificial Intelligence, having being used with slightly different meanings in areas such as planning, problem resolution, decision theory and formal agency theories. Commonly, goals in Agent Programming are considered the same as, or a special kind of, desires. More often then not, goals are taken to be a (logically) consistent subset of the agents desires.

Hindriks et al. (2001) rejects this interpretation by adopting a non-monotonic interpretation for agents goals. Goals in their framework may be viewed an hybrid of desire and intention and, in fact, their agents may pursue jointly inconsistent goals - thus their theory does not conform to Bratman (1999)'s requirements. In their proposal of the GOAL programming language, the authors adopt a declarative interpretation of goals for which a goal is a non-achieved desire held by the agent. An agent, in Hindriks et al. (2001)'s work, is a pair $\langle \sigma, \gamma \rangle$ of sets of propositional formulas, namely the belief base and the goal base of the agent, such that $\gamma \cap Cn(\sigma) = \emptyset$[1], i.e. there is no goal in γ that is already believed by the agent. An agent has a goal that φ (desires that φ), if φ is not believed to hold in the current state of affairs and φ is the consequence of the realization of some goal explicitly held by the agent, i.e. some $\psi \in \gamma$. In another way:

$$\langle \sigma, \gamma \rangle \vDash G\varphi \text{ iff } \sigma \nvDash \varphi \wedge \exists \psi \in \gamma . \psi \vDash \varphi$$

Notice that, by requiring that a formula to be a goal may not be currently believed, i.e. $\sigma \nvDash \varphi$ in the formula above, Hindriks et al. (2001) proposes a non-normal modal logic of goals. Goals and beliefs are connected to intentional action in their framework by means of what they call a conditional action. These conditional actions are formulas of the form $\varphi \rightarrow do(a)$, where φ is a conjunction of $B\xi$ and $G\xi$ formulas and a is a basic action symbol[2]. When an agent $\langle \sigma, \gamma \rangle$ satisfies φ, i.e. $\langle \sigma, \gamma \rangle \vDash \varphi$, the conditional action $\varphi \rightarrow do(a)$ is said to be enabled. The agent must then select one among the enabled conditional actions to execute.

In Hindriks et al. (2001)'s framework, intentions are not explicitly represented as a hierarchical plan, as postulated by Bratman (1999). In their modelling, thus, the agent has only

[1] $Cn(.)$ denotes the (tarskian) consequential closure of propositional logic.

[2] Basic action symbols in this scenario may represent both ontic actions as well as mental actions such as 'adopt a goal ψ.'

unstructured goals that are refined by means of conditional actions, that act as means-end rules.

Van Riemsdijk et al. (2009) expands on the work of Hindriks et al. (2001) giving their notion of goal a default interpretation, clearly inspired by the work of Thomason (2000).To provide this interpretation for the notion of goals, their definition of agent mental state acquire a model theory based on that of default theories - based in extension/answer sets. We believe, however, that this model theory based on extension gives little intuition on the nature of the mental states encoded in their logic or how these models relate to the now classic proposals for logical encoding of mental attitudes, such as that of Cohen and Levesque (1990) and Rao and Georgeff (1998). We believe that a model theory based in Kripke models gives a solution for both problems.

Perhaps the work most related to ours in spirit is that of Hindriks and Meyer (2009). They propose a dynamic logic for agents and show that this logic can be understood as a verification logic, i.e. it has an equivalent state-based semantics established by means of a structural operational semantics. In this logic, the authors encode the attitudes of knowledge and declarative goals, as well as plans and plan adoption rules. The main difference of their approach to ours is that the authors choose to work in a framework closely related to situation calculus. The mental actions involved in decision making and in mental change are, thus, only implicitly defined by means of their encoding of plan adoption, while the inclusion of such actions in the representational language is exactly the main advantage advocated by us. In some sense, our work can be seen as a generalization of the work of Hindriks and Meyer (2009), since by employing Dynamic Preference Logic, the equivalence they seek between operational semantics and declarative semantics can be automatically achieved by the results of Liu (2011) - c.f. Chapters 4 and 5.

Other work have also been proposed for studying the declarative interpretation of intentions in concrete agent programming languages with limited success, in our opinion.

Wobcke (2004), for example, analyses the formal semantics of agent programming languages inspired by the PRS architecture (GEORGEFF; LANSKY, 1987) that does not possess declarative goals, only procedural ones. In this work, the author proposes the Agent Dynamic Logic (ADL), a dynamic logic for the representation of mental attitudes as encoded in these languages. We believe the limitation to procedural goals curbs the understanding of the theory of intentions behind these languages, since we cannot establish a general theory integrating agent's prospective (declarative by nature) intentions with their intentions with which (procedural by nature). Also, by ignoring declarative intentions, his work is only applicable to a subset of the programming languages based on the PRS architecture, excluding the APL family (DASTANI;

VAN RIEMSDIJK; MEYER, 2005; DASTANI, 2008), for example.

On the other way, Bordini and Moreira (2004) present a declarative interpretation of BDI attitudes based on the actual implementation of these concepts in a concrete agent programming language. The aim of their work is to analyse Rao and Georgeff (1998)'s asymmetry properties of mental attitudes encoded in the formal semantics of the language AgentSpeak (RAO, 1996). What is shown in their investigation is that, due to several expressive restrictions in the language, the procedural encoding of mental attitudes in some (early) agent programming languages is very far from the declarative concepts in which they are based.

As such, this methodology of starting from the agent programming language to create a theory of practical reasoning for Agent Programming does not seem fruitful for us. This is because the encoding of mental attitudes in agent programming languages is inevitably influenced by implementation decisions that do not concern the foundational theory.

In this thesis, we will adopt Cohen and Levesque (1990) desiderata, as a compact description of the main points of Bratman (1999)'s requirements for intentions. Regarding the declarative semantics of motivational attitudes, such as desires, we wish to give a Kripke model-based semantics that is related to that of van Riemsdijk et al. (2009), since we believe their work is the most faithful encoding of the notion of goal (or desire) to what is implemented in Agent Programming. This decision also takes into consideration the connection of that work to defeasible reasoning, providing thus a more cognitively sound encoding of desires.

3.2 Intention reconsideration

Largely, the study of changes in intentions are restricted to the now canonical properties introduced by Cohen and Levesque (1990) and Rao and Georgeff (1991) known as Commitment Strategies.

Cohen and Levesque (1990), in their proposal for a logic of intentions, encode the notion of intention in such a way that it satisfies the property, later dubbed as *single-minded commitment*, stated bellow:

> ***Single-minded commitment****: If an agent intends to A, she will maintain such intention until either she accomplishes A or believes it is impossible to A-ing.*

By using intentions as a primitive concept in their logic and not a derived one, as done by Cohen and Levesque (1990), Rao and Georgeff (1991) are able to encode different behaviours for the dynamic of intentions, which they call commitment strategies. Apart from single-minded

commitment, the authors define two extra properties an agent can satisfy in regard to her intentions:

> ***Blind commitment****: If an agent intends to A, she will maintain such intention until she accomplishes A.*

and

> ***Open-minded commitment****: If an agent intends to A, she will maintain such intention until either she accomplishes A, believes it is impossible to A-ing or no longer desires A.*

Rao and Georgeff (1995b) study the problem of intention maintenance and reconsideration in the temporal logic BDI-CTL. The authors start their investigation by showing that a simple axiom stating that an agent will maintain an intention as long as she believes it to be possible is not semantically desirable. This is because, while the agent needs not to intend all the believed consequences of her intentions, she does need to intend the logical consequences of her intentions, since the intention modality is normal in their framework. As such, if an agent intends to get a beer, she would also intend to get a beer or go to the circus - by implication of the disjunction. If all possible intentions are to be maintained, if the agent comes to believe she can no longer have a beer, since the second intentions is yet possible, she would have to intend to go to the circus.

To solve this problem and to express the desirable properties for intention maintenance, the authors argue for the need to extend their logic to include the notion of explicitly intending (also explicitly believing and explicitly desiring), i. e. by adding modal operators to intention in the sense of Levesque (1984). By encoding the notion of changes in the beliefs of the agent - by means of what they call Belief Revision Function -, the authors solve the problem of preservation of undesirable intentions requiring that only explicitly intended states are maintained. As a theory of intention reconsideration, this work poses very few conditions for an intention to be reconsidered, namely when it is not believed to be possible. The more general problem of intention dynamics, however, is left as future work by the authors.

Apart from the focus on commitment strategies, a considerable portion on the literature about the dynamics of intentions focus on the reconsideration of plans (KONOLIGE; POLLACK, 1993; VELOSO; POLLACK; COX, 1998; VAN DER HOEK; JAMROGA; WOOLDRIDGE, 2007; ICARD; PACUIT; SHOHAM, 2010). Using a plan-based representation of intentions, they specify the causal forces to adopt and abandon an intention. The main innovation in this approach, in our opinion, is that differently than previous work, instead of an

intrinsic property of intentions, the persistence of intentions can be analysed as a side-effect of the beliefs about the plans. Particularly, Wobcke (1996) claims that the persistence of intention is nothing but a side-effect of the principle of minimal change in the beliefs of the agent.

Wobcke (1996) presents a theory for plan and intention reconsideration based on belief revision. In his theory, the agent's belief base store both the structure of plans of the agent - as means-end beliefs - and the currently held intentions of the agent. In his theory, thus, an agent belief base is something like the one depicted in Figure 3.1 for an agent wanting to alleviate the heat of a warm day by drinking a cold beer.

Figure 3.1 – A belief base of an agent as described by Wobcke.

$$B(warm(today)) \rightarrow I(get(beer))$$
$$B(warm(today))$$
$$B(cold(beer))$$
$$I(get(beer)) \wedge B(cold(beer)) \rightarrow I(go(kitchen))$$
$$I(get(beer)) \wedge \neg B(cold(beer)) \rightarrow I(go(bar))$$

Source: the author.

Given that intentions are subjected to Bratman (1999)'s strong consistency requirements, a change in the agent's belief base automatically trigger a change in her plans and intentions. Going back to the example of Figure 3.1, if the agent comes to discover that today is not indeed a warm day - she just probably forgot to turn off the heat on the house -, she will automatically revise her intention to get a beer.

Notice that Wobcke's work has two major drawbacks. On one hand, as a theory of practical reasoning, this approach is not advantageous since, by representing intention and plans in the agent's belief base, the agent must continuously perform theoretical reasoning on her beliefs to decide what to do. Well, this need for constant reconsideration is exactly what Bratman, Israel and Pollack (1988) emphasize as one of the key pragmatical functions of intentions.

On the other hand, as a theory of intention dynamics, his approach is limited in the sense that it does not provide reasons for giving up or reconsidering one's intention, such as reconsideration strategies, unless in the specific case of the supporting beliefs of an intention having been dropped. In fact, his approach does not provide reasons to drop even those intentions which the agent does not believe to be achievable - i.e. for which the agent does not have a plan to achieve them.

Van der Hoek, Jamroga and Wooldridge (2007) construct an algorithmic theory of plan reconsideration. The main problem treated by the authors is how to (possibly) change the plans the agent has adopted in order to achieve a certain intention given a change in the agent's beliefs.

An intention, in their framework, is always maintained as long as is has not been achieved and there is a possible plan to achieve it. In other words, these authors propose an algorithm to implement single-minded agents.

Another interesting approach to intention reconsideration is the work of Icard, Pacuit and Shoham (2010). These authors, inspired by Shoham (2009)'s database perspective, propose a formal theory of intentions and beliefs and connected the dynamics of such notions to the work in AGM's Belief Revision Theory. Following the database perspective, an intention is a set of temporally annotated actions that must be performed by the agent. The reconsideration aspect of this theory concerns the maintenance of the underlying guiding principles of their theory - belief-intention consistency and coherency - that must be preserved under AGM's revision of the agent's beliefs, similar to Wobcke (1996).

Recently, however, Van Zee et al. (2015b) showed that the logic presented by Icard, Pacuit and Shoham (2010) is not sound and, in fact, such logic is not compact. As such, there is no finite axiomatization for it. They propose a change in Icard et al.'s language by explicitly annotating each propositional symbol and modality with the temporal point to which it refers, constructing in such a way an hybrid doxastic dynamic logic. They also focus on redefining Icard, Pacuit and Shoham (2010)'s notions of belief-intention consistency and coherency, which they argue are too permissive in the original work.

In later work, Van Zee et al. (2015a) propose the reconstruction of the connection between AGM's Belief Revision to the dynamics of belief and action in their framework. Exploring the limited expressibility of their language, provided by both the requirement of deterministic actions and the explicit temporal reference, they apply Katsuno and Mendelzon (1991)'s codification of propositional models to provide a revision operator for their logic satisfying AGM's conditions.

These approaches, in a sense, try to encode the case argued for reconsideration only when it violates the strong consistency requirements. As such, they focus on how to restore mental state coherency in the face of belief change.

Thangarajah, Padgham and Harland (2002), on the other hand, use an explicit representation of preference among goals and rules to represent policies for goal and plan adoption. This work extends Rao and Georgeff's commitment strategies proposing various requirements for plan and goal adoption and reconsideration. The authors propose a set of alternative rules,

such as R_4 below, to specify the behaviour of the agent.

$$R_4: \quad ExStep(\alpha, P_1, \phi) \wedge ExStep(\beta, P_2, \phi') \wedge \\ Con(\phi, \phi') \wedge AltPlan(\alpha, P_1) \wedge \neg AltPlan(\beta, P_2) \rightarrow Pref(\phi', \phi)$$

In the rule above, presented by Thangarajah, Padgham and Harland (2002, p.4), it is stated that if ϕ is a step in a plan P_1 for an intention α (denoted by $ExStep(\alpha, P_1, \phi)$ in R_4 above) such that ϕ conflicts with a step ϕ' (denoted by $Con(\phi, \phi')$ in R_4 above) in a plan P_2 for an intention β (denoted by $ExStep(\beta, P_2, \phi')$ in R_4 above) and there is an alternative plan to P_1 for intention α but not to P_2 for β, then the agent should prefer to execute ϕ' than ϕ.

In a sense, Thangarajah, Padgham and Harland (2002) propose a framework to specify and implement some form of Bratman (1999)'s reconsideration policies. The authors also evaluate empirically the efficiency of some of the proposed policies in the agent programming language JACK (HOWDEN et al., 2001).

On the study of reconsideration policies and their empirical evaluation some work has shown how different policies for reconsideration can impact the behaviour of an agent according to different characteristics of the environment.

Kinny and George (1991) present the first empirical evaluation of how different commitment policies impact the agent's effectiveness, based on environmental characteristics, such as cost of planning and how often the environment change. They show that, given the costs associated with deliberation, if an agent inhabits an environment where change is not frequent, blind commitment is a advantageous policy. On the other hand, as the environment becomes more dynamic, the ability for one agent to re-plan and reconsider becomes an advantageous feature.

For the authors, the following events lead to reconsideration: the trigger event of an intention has ceased to be true; new events have been detected; a more relevant event, or with higher priority, is detected. Their empirical study shows that the first and the last conditions have the most impact in rational optimality of the agent.

Wooldridge and Parsons (1999) present a utilitarian framework to reason about reconsideration. The authors show how a decision theoretic approach can be employed to guide high-order reasoning, i.e. when to deliberate, when to deliberate whether to deliberate, and so forth. In a sense, they propose an explanation of Kinny and George (1991)'s results by presenting a formal framework to study reconsideration policies.

Schut (2002) generalizes the study of Wooldridge and Parsons (1999) by incorporating several features in their model for second-order reconsideration, i.e. when to deliberate

whether to deliberate. As for Wooldridge and Parsons (1999), his methods are not based on the consistency and coherence forces that constrain the agent's mental state, but rather based on a utilitarian framework they propose to evaluate the gain associated with acting or reconsidering. Finally, this author shows how second-order deliberation can be further generalized as a Markovian Decision Process and provides algorithms for the its integration inside a BDI agent interpreter, *a la* Rao and Georgeff (1995a).

Grant et al. (2010) propose a theory of intention revision, based on a utilitarian approach to intention reconsideration. They present a set of postulates that, together with the well-known AGM postulates (ALCHOURRÓN; GÄRDENFORS; MAKINSON, 1985) for belief, describes "minimal change" principles for an agent mental state.

The authors propose a formal framework to describe an agent mental state, giving a set of rationality postulates for the relation between the agent's cognitive structures, i.e. her mental attitudes, as well as a set of postulates governing the change in each mental attitude. Most important, regarding the changes in intentions, the authors require that the changes in the agent's intentions are (set-theoretically) minimal.

While an interesting proposal, in the sense that it provides a holistic theory for mental state dynamics, we believe their approach has some important drawbacks for the perspective of a theory for Agent Programming. Firstly, as in the case of AGM's Belief Revision approach, their handling of mental state dynamics is extra-semantic, in the sense it cannot be expressed inside the logic constructed to reason about the agent's minds, only by a outside observer. Secondly, while their approach highlights the necessary conditions for a change of mind to be rational - based on minimality of change and value maximization - it does very little to clarify the mechanisms governing the change of mind. In a sense, Grant et al. (2010) propose a normative set of conditions for one's change of mind to be rational, but not how one is supposed to change her mind.

From another perspective, Meyer, Van der Hoek and van Linder (1999), and later Van Riemsdijk et al. (2005), model commitment to an intention as a conscious action performed by the agent. In this perspective, the agent given her beliefs chooses to commit to a given (acceptable) goal and once committed, maintain this intention until completion, impossibility or until she decides to revoke her commitment.

Meyer, Van der Hoek and van Linder (1999) presents the KARO framework, a dynamic logic for agent specification representing agents knowledges, abilities, results and obligations. In this logic, the authors model different motivational attitudes, such as wishes and intentions. The interesting insight provided by this approach is that committing and uncommitting to an

intention is modelled as an explicit mental action of the agent. As such, the reconsideration of an intention is not a consequence of the triggering event that led the agent to balance her reasons, but the conscious choice of the agent to forfeit commitment and consider if there are better options to achieve a given desire.

Van Riemsdijk et al. (2005) define mechanisms for change in the agents goals based on agents beliefs. The authors show how to define different strategies for goal adoption and dropping based on an operational semantics for a BDI agent programming language similar to 3APL (DASTANI; VAN RIEMSDIJK; MEYER, 2005). The authors expand on the already known mechanisms for specifying goal change by exploring the use of non-monotonic rules to define goal adoption and goal dropping.

An important (practical) aspect of these goal generation rules is the mechanism of goal reconsideration based on failure conditions. A failure condition is a non-monotonic rule $\beta \rightarrow_G^- \varphi$ specifying that if an agent believes in β, then she has a reason to drop goal φ. We believe the use of goal generation rules is an interesting mechanism for the specification and implementation of an agent's reconsideration strategies.

In our work, in Chapter 5, we will adopt the idea of mental change as actions performed by the agent, as proposed by Meyer, Van der Hoek and van Linder (1999) in their KARO framework. We believe this approach to be more general, in the sense that by means of representing explicitly the mental actions involved in mental change, one can easily specify different policies for intention and plan reconsideration discussed before, while guaranteeing the maintenance of basic semantic properties - as Bratman (1999)'s requirements for mental coherency and Mintoff (2004)'s reconsideration principles presented in Chapter 2.

3.3 Belief and mental attitude dynamics

The dynamics of mental attitudes is the study of *how* and *why* a given agent comes to change her positions about the environment she inhabits. From all the mental attitudes studied in the philosophical literature, the dynamics of beliefs has received the most attention in the literature, for its central role in Epistemology and Philosophy of Science.

Belief revision is the area that studies how a doxastic agent change her beliefs in face of new information - which may possibly conflict with currently held beliefs. Currently, the most influential model for belief revision is the so-called AGM paradigm, named after the authors of its seminal paper (ALCHOURRÓN; GÄRDENFORS; MAKINSON, 1985). Although the authors approach and philosophical hypothesis have been questioned in the literature (ROTT,

2000; KATSUNO; MENDELZON, 1992; HANSSON, 1992; POLLOCK, 2001), it is unquestionable that it has brought profound developments for the problem of belief dynamics, influencing works on areas such as Computer Science, Artificial Intelligence and Philosophy.

Segerberg (2001) proposed the codification of AGM revision operations within a dynamic modal logic. That change is important because it extends the expressibility of the logic and allow one to analyse the effects of introspection, and other related phenomena, in the logic of belief change. However, DDL has still some limitations such as the impossibility of representing the nature (or structure) of the new acquired information and a have cumbersome semantics. To overcome this problem, Van Benthem (2007) proposed a codification for Belief Revision operations inside Dynamic Epistemic Logic (DEL), which was further developed by Baltag and Smets (2008).

In the following, we will revisit the AGM tradition of Belief Revision and, further, introduce the main ideas of the Dynamic Epistemic Logic approach, providing its connections to the study of belief change *a la* AGM.

3.3.1 AGM Belief Revision

The seminal paper of Alchourrón, Gärdenfors and Makinson (1985) introduced the AGM paradigm for belief revision. The AGM approach focused on defining the requirements for rational changes of the agents beliefs, which the authors claim to encode the Quine (1951)'s requirement for minimal mutilation of the web of beliefs, translated by the authors into minimal change.

The AGM approach focus on the revision sets of beliefs (or theories). A belief set K from a given Logic $\mathscr{L}$ consists of a closed set over the consequence operator Cn of $\mathscr{L}$, i.e., K is a set of sentences from the language of $\mathscr{L}$ satisfying $K = Cn(K)$. Over such belief set, some operations are defined: expansion, contraction and revision.

The expansion operation consists of adding new beliefs to an existing belief set, the contraction consists of removing from the set a particular belief and revision consists in adding new belief to the set consistently, i.e. add new beliefs to the set so that the result of the operation is a consistent belief set. Among the three operations, only the expansion $+$ can be univocally defined (ALCHOURRÓN; GÄRDENFORS; MAKINSON, 1985): $K + \varphi = Cn(K \cup \{\varphi\})$.

For other operations, the authors provide a set of postulates that they deem to characterize the appropriate behaviour for them. These postulates, commonly called AGM postulates or Gärdenfors postulates, do not completely characterize an operator, but, rather, define a class of

operators on sets of beliefs that can be used to contract (or revise) such sets.

Let K be a belief set, i.e. a theory over the logic $\mathscr{L} = \langle L, Cn \rangle$, with Cn a consequence operator[3] on the language of $\mathscr{L}$, a rational contraction operation must satisfy:

(C-1) $K - \alpha = Cn(K - \alpha$

(C-2) If $\alpha \notin Cn(\emptyset)$, then $\alpha \notin K - \alpha$

(C-3) $K - \alpha \subseteq K$

(C-4) If $\alpha \notin K$, then $K - \alpha = K$

(C-5) If $Cn(\alpha) = Cn(\beta)$, then $K - \alpha = K - \beta$

(C-6) $(K - \alpha) + \alpha = K$

The postulate C-1 provides that the contraction still results in a belief set. The postulated C-2 , usually named Success, ensures that if the formula α one wants to extract from the set of beliefs is not a tautology - and therefore necessarily true - then, after the contraction of α from the belief set, this formula is no longer believed by the agent . C-3 ensures that any new formula, not previously believed by the agent, will not be included in her set of beliefs after contraction. C-4 says that , if the belief set does not prove the formula to be removed, so no changes in the agent's beliefs are needed. C-5, commonly called extensionality principle, ensures that contraction takes into account the semantics of the formula and not its syntactic structure, and finally, C-6, commonly named recovery, ensures that only formulas related to the belief being removed will be removed from the set of beliefs.

Similarly, the postulates for revision are presented:

(R-1) $K * \alpha = Cn(K * \alpha)$.

(R-2) $\alpha \in K * \alpha$.

(R-3) $K * \alpha \subseteq K + \alpha$.

(R-4) If $\perp \notin K + \alpha$, then $K * \alpha = K + \alpha$.

(R-5) If $Cn(\alpha) = Cn(\beta)$, then $K * \alpha = K * \beta$.

(R-6) If α is consistent, then so is $K * \alpha$.

These three operations are interconnected by the properties known as Levi and Harper identities. Through them, it is possible to define a revision operation, by means of the contraction operation and expansion:

$$K * \alpha = (K - \neg\alpha) + \alpha$$

[3]Notice that, in the original article, AGM requires a supra-classical compact tarskian logic satisfying disjunction introduction rule, although it is usually taken as either classical propositional or predicate logic.

and, reciprocally, a contraction operator by means of revision and expansion:

$$K - \alpha = K \cap (K * \neg\alpha).$$

Many criticism arose against AGM's work contesting their philosophical adequacy, the conceptual underpinings of their theory or proposing new notions of epistemic and doxastic change (ROTT, 2000; KATSUNO; MENDELZON, 1992; HANSSON, 1992; POLLOCK, 2001). Although it is wildly recognized that the AGM approach relies on severe idealization of agents and their capabilities, it is hard to argue with the applicability and influence of their work on many areas such as Philosophy, Logic, Artificial Intelligence and Computer Science.

Particularly, Hansson (1992) criticises the use of deductively closed set of formulas and to the AGM postulates, providing examples for which the structure of the beliefs of an agent may influence the change, not just their meaning. This author proceeds to construct a different notion of belief revision which rely on the structure of the information believed by the agent - which he calls Belief Base Revision.

Aiming for a change operation that relies on a explicit codification of an agent's doxastic commitments, not possible with belief bases, Williams (1994) proposes a relational model for syntactic-based Belief Revision, similar to Gärdenfors and Makinson (1988)'s entrenchment relations, called ensconcement-based revisions. This model is shown to describe a subclass of Hansson's Belief Base revision functions (FERMÉ; KREVNERIS; REIS, 2008).

Williams define an ensconcement as a set of formulae Γ together with a total pre-order $\preceq$ over Γ satisfying the following conditions:

- For all nontautological $\beta \in \Gamma$, $\{\alpha \in \Gamma : \beta \prec \alpha\} \nvdash \beta$
- For all $\beta \in \Gamma$, $\alpha \preceq \beta$ for all $\alpha \in \Gamma$ if and only if $\vdash \beta$

Using esconscement relations, the author defines a syntax-based revision operation. This definition however cannot be used iteratively since the resulting product of the operation is a set of formulas is a set, not an ensconcement. Investigating iterated belief change in this framework, the author later proposes a construction, based on Ordinal Conditional Functions (SPOHN, 1988). This constructive approach, however, relies heavily on numerical codification of epistemic certainty and is strongly criticized by Rott (98) for unexpected behaviours it presents.

Drawing from the iterated belief revision methods by Darwiche and Pearl (1997), Jin and Thielscher (2007) propose a new postulate - along with a OCF-based construction - to overcome to limitations of Darwiche and Pearl's work. In later work, Jin, Thielscher and Zhang (2007)

show how this new operator may be used to merge two different belief bases - represented as OCFs.

While most the philosophical and pragmatical aspects of Belief Revision in the AGM tradition have been extensively studied, this approach remains still essentially as a extra-logical exploration of the dynamics of belief. Only recently these dynamic aspects of Belief - following the AGM tradition - were embedded in a logical framework that allowed the study of the interplay between the now static phenomena of Belief and their dynamics. These investigations, such as Segerberg (2001)'s Dynamic Doxastic Logic , provide us more flexible tools to understand problems and limiting cases that previous theories cannot express, e.g. higher-order revision and Moorean sentences. In the following we will focus on one such approaches based on the Dynamic Epistemic Logic tradition.

3.3.2 DEL and Belief Change

While changes in mental attitudes have been a well studied topic in the literature, the integration of such operations within the logics of beliefs, obligations, desires and others is a somewhat recent development. To our knowledge, the work of Segerberg (1999) is the first to propose the integration of dynamic logic-like operations within the logic of Beliefs and Knowledge to represent the doxastic changes as studied in AGM tradition (ALCHOURRÓN; GÄRDENFORS; MAKINSON, 1985).

This shift from extra-logical characterization of changes in the agents attitudes to their integration within the representation language has important expressibility consequences. It allows, for example, the study of dynamic phenomena not representable in axiomatic approach of the AGM framework. This is the case, for example, of the well-known Moore sentences in Epistemology (MOORE, 1993).

Recently, inspired by Van Benthem and the Dutch School on the "dynamic turn" in Logic (VAN BENTHEM, 1996), several dynamic logics for information change and dynamics of mental attitudes have been proposed (VAN BENTHEM, 2007; BALTAG; SMETS, 2008; VAN BENTHEM; GIRARD; ROY, 2009; LIU, 2011; VAN BENTHEM; PACUIT; ROY, 2011).

Inspired by the connection between belief revision policies and transformations in priority structures, presented by Rott (2006), Van Benthem (2007) embeds some belief change operations in the framework of Dynamic Epistemic Logic. Baltag and Smets (2008) consolidate this connection by providing a semantic codification of different epistemic and doxastic attitudes, such as Safe Belief, Knowledge, Conditional Belief, etc. and providing axiomatiza-

tion for both the static and dynamic parts of this language. Finally, Baltag, Fiutek and Smets (2014) show that the unlimited Dynamic Doxastic Logic - an extension of the original logic of Segerberg (1999) - is expressively equivalent to Dynamic Epistemic Logic, and thus, just another formalism to express the same phenomena.

Studying the logic of preferences, Girard (2008) and Van Benthem and Liu (2007) generalize the results of Baltag and Smets (2008), presenting a logic for preferences and order, which was used to encode several different notions, such as Conditional Preferences, Beliefs, Obligations, Contrary-to-Duty reasoning, etc. These works extend the Dynamic Epistemic Logic approach to investigate change in several mental attitudes by means of preferential models, similar to the ones used in the area of non-monotonic reasoning. In that line, the work of Van Benthem and Liu (2007) can be seen as the first programmatic attempt to study of information change using PDL as a methodological tool to acquire reduction axioms for the resulting logic. This approach was later developed by Liu in a series of papers which culminated in her 2011 book (LIU, 2011).

The work of Liu (2011) provides the connection between preference models, generally used to encode mental attitudes such as beliefs (BALTAG; SMETS, 2008), obligations (VAN BENTHEM; GROSSI; LIU, 2014) etc., and syntactic structures providing justifications - known in their work as priority graphs. The author further shows that several change operations over possible-worlds models, as defined in the Dynamic Epistemic Logic tradition, can be equivalently represented as syntactic transformations on these priority graphs.

This work is particularly important to our study because we believe this connection between possible worlds models and syntactic codifications is exactly the kind of result that a study of the semantic codification of mental attitudes in agent programming languages seeks to establish. As such, if one can explore these syntactic representations to implement agent programming languages, the connection between the programming language semantics - as usually described by means of operational semantics (PLOTKIN, 2004) or interpreters - and the declarative semantics provided by a BDI logic is an immediate result. As a result, we can reason about an agent program execution by means of the declarative interpretation of the program on the associated logic.

Preference models lie at the core of the formalization for several related notions, such as non-monotonic reasoning, obligations, goals, beliefs and preferences. Logics for these concepts are now well-established in the literature. Particularly, the Dynamic Preference Logic introduced by Girard (2008), a logic in the family of Dynamic Epistemic Logics, has shown great flexibility to encode these notions, as well as actions representing change policies in an

agent's mental state. Therefore, this logic has the potential to become a common framework to study several related notions in Philosophy and Artificial Intelligence.

3.4 Summary of the chapter

In this chapter, we presented the State of the Art in the formalization of mental attitudes and the formal treatment dedicated to some phenomena involving the attitudes, specially the problem of Intention Reconsideration and mental state change.

We presented, in Section 3.1, some of the most important formal representation of mental attitudes in the area of Artificial Intelligence, noting the conceptual differences for the terms between the areas. In that section, we gave a particular focus to formalization of intentions, but also discussed the notion of desires and goals.

Based on the discussion of Section 3.1, we choose to adopt Cohen and Levesque (1990)'s desiderata, as a compact description of the main points of Bratman (1999)'s requirements for intentions. We will establish a declarative semantics for intentions (as well as desires) based on Kripke semantics, which we believe to express more clearly the connection between the formalization and the philosophical foundation we chose. While Kripke models have been criticized as a semantic tool for their distance to the computational representations commonly used to represent mental attitudes in agent programming languages, we believe that exploring the connection established by Liu (2011) between Kripke models and some syntactic structures, we can overcome this semantical gap and provide both a well-studied declarative semantics based on Kripke models and a practical way to implement agent programming languages.

In Section 3.2 we discussed formal models for intention reconsideration. In that section we discussed the many proposals of reconsideration policies and their evaluation in the literature of Agent Programming. In our work, in Chapter 5, we will adopt the idea of mental change as actions performed by the agent, as proposed by Meyer, Van der Hoek and van Linder (1999) in their KARO framework.

We believe this approach to be more general, in the sense that by means of representing explicitly the mental actions involved in mental change, one can easily specify different policies for intention and plan reconsideration discussed before, while guaranteeing the maintenance of basic semantic properties - as Bratman (1999)'s requirements for mental coherency and Mintoff (2004)'s reconsideration principles presented in chapter 2.

Finally, in Section 3.3, we presented some the basic notions of Dynamic Epistemic Logic, the logical framework we will adopt in this work. In this discussion, we defend our

choice of representing mental changing actions as part of the representation language, as well as point out the methodology we will adopt in the construction of the formal base for our work: namely, the program of studying information change using PDL sketched by Van Benthem and Liu (2007).

4 DYNAMIC PREFERENCE LOGIC

Preference Logic (or Order Logic as named by Girard (2008)) is a modal logic complete for the class of transitive and reflexive frames. It has been applied to model a plethora of phenomena in Deontic Logic (VAN BENTHEM; GROSSI; LIU, 2014), Logics of Preference (BOUTILIER, 1994b; LANG; VAN DER TORRE; WEYDERT, 2003), Logics of Belief (BOUTILIER, 1994a; BALTAG; SMETS, 2008), and also in Non-monotonic reasoning (KRAUS; LEHMANN; MAGIDOR, 1990).

Dynamic Preference Logic (DPL) (GIRARD, 2008; VAN BENTHEM; GIRARD; ROY, 2009) is the result of "dynamifying" Preference Logic, i.e. extending it with dynamic modalities - usually represented by programs in Propositional Dynamic Logic (PDL). This logic is interesting for its expressibility, allowing the study of dynamic phenomena of attitudes such as Beliefs, Obligations, Preferences etc. For that, Dynamic Preference Logic has the potential to become a common framework to study different but correlated notions in the areas of Epistemology, Deontic Logic and Decision Theory. It is one of many dynamic logics proposed in the tradition of Dynamic Epistemic Logic, sharing similar methodologies and tools.

In this chapter, we present the language and semantics of Dynamic Preference Logic. Particularly important in this exposition is the connection between the preference models used to define the semantics of that logic and the syntactic structure of priority graphs defined by Liu (2011). Our main interest in this connection is that priority graphs are computational-friendly structures that may be used to reason about agent's preferences. As such, in Chapter 6, we will adopt these structures as a mean to both implement agent programming languages and, by exploring the connection between priority graphs and preference models, to obtain the relation between agent programs and logical models for agents.

Firstly, in Section 4.1, we present the "static" Preference Logic based in the work of Girard (2008). In order to introduce some important dynamic operators in the logic, namely contractions, in Section 4.2 we change Girard (2008)'s semantics to require further that preference models are well-founded and we provide a sound and complete axiomatization of the logic. In Section 4.3, we explore the connection established by Liu (2011) between preference models used in the logic and syntactic structures, known as priority graphs, which we believe to be computationally-friendly structures to reason about preferences. Later, in Sections 4.4 and 4.5, we "dynamify" this logic by further introducing dynamic modalities representing standard belief change operations such as revisions and contractions.

While in Section 4.4 we merely present some operations studied in the literature in

the context of Dynamic Preference Logic, in Section 4.5, we propose contraction operators for Dynamic Preference Logic, based on the study of iterated belief contraction operations of Ramachandran, Nayak and Orgun (2012). As far as we know, ours is the first proposal of codification of such operations in Preference Logic. Additionally, we show that some of these contraction operations cannot be defined by means of priority graphs. This result is a negative answer to the question posed by Liu (2011) about whether any PDL-definable operation closed for the class of preference models is harmonic, i.e. can be equivalently represented by transformations in priority graphs. We point out that, to our knowledge, this is the first work to investigate harmony properties for contraction operators in the literature of Dynamic Epistemic Logic.

Finally, in Section 4.6, we show that a particular class of preference models, called broad models, is ideal in the sense that these models satisfy properties that allow us to reason about preferences and preference change by means of the priority graphs we discuss in Section 4.3.

The logic and operations presented in this chapter compose the base formalism we explore in our work. In Chapter 5, we will apply the results presented here to create a logic to specify BDI agents from the point of view of agent programming. Further, in Chapter 6, we explore the connection between priority graphs and preference models, discussed in Section 4.3, to obtain declarative interpretations for agent programs.

4.1 A logic of static preferences

Let's first introduce the language of Preference Logic.

Definition 4.1 *Let P be a finite set of propositional letters. We define the language $\mathscr{L}_{\leq}(P)$ by the following grammar (where $p \in P$):*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid A\varphi \mid [\leq]\varphi \mid [<]\varphi$$

We will often refer to the language $\mathscr{L}_{\leq}(P)$ simply as $\mathscr{L}_{\leq}$, by supposing the set P is fixed. Also, as customary in the modal logic literature, we will denote the language of propositional formulas, i.e. the language removing all modal formulas from $\mathscr{L}_{\leq}(P)$, by $\mathscr{L}_0(P)$ or simply $\mathscr{L}_0$

Definition 4.2 *A preference model is a tuple $M = \langle W, \leq, v\rangle$ where W is a set of possible worlds, $\leq$ is a a reflexive, transitive relation over W, and $v : P \to 2^W$ a valuation function.*

In such a model, the accessibility relation $\leq$ represents an ordering of the possible worlds according to the preferences of a certain agent. As such, given two possible worlds $w, w' \in W$, we say that w is at least as preferred as w' if, and only if, $w \leq w'$. While we will commonly use the term 'preference relation', we wish to point out that the interpretation for that relation depends on the application of the logic. As such, when using preference logic to encode beliefs, the accessibility relation $\leq$ is commonly referred as a 'plausibility relation' among worlds, denoting which state of affairs the agent beliefs to be more plausible. On the other hand, when using this logic to study deontic phenomena, the same relation is commonly referred as a 'betterness relation.'

The interpretation of the formulas over these models is defined as usual. We will only present the interpretations for the modalities, since the semantics of the propositional connectives is clear. The A modality is an universal modality[1] satisfied iff all worlds in the model satisfy its argument. The $[\leq]$ modality is a box modality on the accessibility order $\leq$. The $[<]$ modality is the strict variant of $[\leq]$. They are interpreted as:

$$\begin{array}{lll} M, w \vDash A\varphi & \text{iff} & \forall w' \in W : M, w' \vDash \varphi \\ M, w \vDash [\leq]\varphi & \text{iff} & \forall w' \in W : w' \leq w \Rightarrow M, w' \vDash \varphi \\ M, w \vDash [<]\varphi & \text{iff} & \forall w' \in W : w' < w \Rightarrow M, w' \vDash \varphi \end{array}$$

As such, the formula $A\varphi$ can be read as '*universally* φ' or '*it is universally true that* φ', while the formulas $[\leq]\varphi$ and $[<]\varphi$ can be read as '*in every situation at least as preferable as the current one,* φ *holds*' and '*in every situation strictly preferable than the current one,* φ *holds*', respectively.

We will refer as $E\varphi$ to the formula $\neg A \neg \varphi$, meaning '*it is possibily true that* φ', and as $\langle\leq\rangle\varphi$ ($\langle<\rangle\varphi$) to the formula $\neg[\leq]\varphi$ ($\neg[<] < \varphi$), meaning '*in a possible situation at least as (strictly more) preferable as the current one,* φ *holds*,' as commonly done in modal logic.

As usual, given a model M and a formula φ, we use the notation $[\![\varphi]\!]_M$ to denote the set of all the worlds in M satisfying φ. When it is clear to which model we are referring to, we will denote the same set by $[\![\varphi]\!]$. Also, given a set of worlds $[\![\varphi]\!]$ and a (pre-)order $\leq$, we will denote the minimal elements of $[\![\varphi]\!]$, according to the strict part $<$ of the relation $\leq$, by the notation $Min_{\leq}[\![\varphi]\!]$. This corresponds to the notion of 'most preferred worlds satisfying φ' in the model.

A complete axiomatization for the above logic has been provided by Girard (2008). Since we will change the semantics to require a further restriction of the preference models used

[1] In this work, we understand the worlds as epistemically possible worlds, not metaphysically possible. While formally this difference is irrelevant, philosophically it is of importance. Particularly, when we define the notion of knowledge in Chapter 5.

in our work, we will delay the presentation of an axiomatization for the logic until Section 4.2.

As the concept of most preferred worlds satisfying a given formula φ will be of great use in modelling some interesting phenomena in this logic, we define a formula encompassing this exact concept.

Definition 4.3 *We define the formula $\mu\varphi \equiv \varphi \wedge \neg\langle<\rangle\varphi$, that is satisfied by exactly the most preferred worlds satisfying φ, i.e. $[\![\mu\varphi]\!]_M = Min_{\leq}[\![\varphi]\!]_M$.*

An interesting kind of formula can thus be defined in this logic, namely *conditional preference*. A conditional preference is a dyadic modality $C(\psi|\varphi)$ expressing the information that '*in the most preferred φ-worlds, ψ holds*.':

Definition 4.4 *We call the conditional preference of ψ given φ the formula:*

$$C(\psi|\varphi) \equiv A(\mu\varphi \rightarrow \psi)$$

Conditional statements such as the above have been expressed by means of dyadic modalities since Chisholm (1963)'s analysis of contrary-to-duty obligations in Deontic Logic. Conditionals are common in planning and practical reasoning, being used, for example, to express dependency relations among the agent's desires.

Notice that, for any preference model $M = \langle W, \leq, v\rangle$ and world $w \in W$, $M, w \vDash C(\psi|\varphi)$ if, and only if, $Min_{\leq}[\![\varphi]\!] \subseteq [\![\psi]\!]$. In other worlds, M only satisfies $C(\psi|\varphi)$ if the minimal, or most preferred, worlds satisfying φ also satisfy ψ, as the intuition of $C(\psi|\varphi)$ required.

These formulas will be of particular importance in Chapter 5, since they will be used to encode the notion of mental attitudes, such as belief and desire. In fact, it has been argued that conditional modalities offer a better representation of some mental attitudes such as obligations (VAN BENTHEM; GROSSI; LIU, 2014), beliefs (BALTAG; SMETS, 2008) and desires (DOYLE; SHOHAM; WELLMAN, 1991).

4.2 Paving the way to a Dynamic Preference Logic

It has been observed by Girard and Rott (2014) that some operations are only well-defined for a special class of preference models. Namely, the condition required on these preference models is the satisfaction of a well-known property in the study of Belief Change: the Lewis Limit Assumption. The Lewis Limit Assumption requires that for any satisfiable formula φ, the set $[\![\varphi]\!]$ of worlds satisfying φ has minimal elements, i.e. $Min_{\leq}[\![\varphi]\!] = [\![\mu\varphi]\!] \neq \emptyset$. This

assumption is equivalent to the requirement that if φ is satisfiable, then there is no infinite descending chain of worlds satisfying φ. A set of important operations that require such condition in order to be well-defined are the contraction operations we study at Section 4.5.

Since Lewis Limit Assumption is intrinsically dependent on the language[2], some authors, such as Baltag and Smets (2008) and Girard and Rott (2014), advocate for a purely semantic restriction on models, i.e. not constrained by any particular language, that entail the limit condition. This restriction is the well-foundedness[3] of the strict part $<$ of the accessibility relation $\leq$. From now on, we will call any preference model with a well-founded strict part $<$ a well-founded preference model.

Until now, however, it was an open problem to provide an axiomatization for Preference Logic restricted to well-founded preference models. Some authors, notably Baltag and Smets (2008) and Girard and Rott (2014), pointed out the relationship between well-foundedness and Löb Axiom, as well studied in the case of Provability Logic (JAPARIDZE; DE JONGH, 1998; VAN BENTHEM, 2006).

Girard and Rott (2014) suggest the addition of this axiom would suffice to guarantee well-foundedness of the models, but cannot prove such conjecture for their logic. In what follows, we will give the proof of such claim. First, we show that any preference model for which the accessibility relation $\leq$ satisfies reflexiveness and transitivity, it satisfies Löb Axiom for a formula φ if, and only if, it satisfies that $[\![\varphi]\!]$ has minimal elements, i.e. it satisfies the Limit Assumption. This result will be a stepping stone to provide, later in this section, the full proof of the soundness and completeness for the axiomatization in regard to well-founded preference models.

For the first proof, we will use an equivalent formulation (W') of Löb Axiom ($W : ([<]\varphi \rightarrow \varphi) \rightarrow [<]\varphi$) with possibilities instead of necessities:

$$\mathbf{W}' : \quad \langle < \rangle \varphi \rightarrow \langle < \rangle (\varphi \wedge \neg \langle < \rangle \varphi)$$

Axiom W' above can be understood as '*if there is a more preferable element satisfying* φ, *then there is a minimal more preferable element satisfying* φ'.

Lemma 4.5 *Let* $M = \langle W, \leq, v \rangle$ *be a reflexive and transitive model and* $\varphi \in \mathscr{L}_{\leq}(P)$, *s.t.* $[\![\varphi]\!]_M \neq \emptyset$. *Then, for any* $w \in W : M, w \vDash \langle < \rangle \varphi \rightarrow \langle < \rangle (\varphi \wedge \neg \langle < \rangle \varphi)$ *iff* $Min_{\leq} [\![\varphi]\!]_M \neq \emptyset$.

[2]Notice it is stated based on the formulas of the language.

[3]A relation $R \subseteq W^2$ is well-founded iff $Min_R S \neq \emptyset$ for all non-empty $S \subseteq W$, i.e. every non-empty subset of W has minimal elements or, equivalently, there are no infinite descending chains of worlds in W.

Proof:

$\Rightarrow$:

Since $[\![\varphi]\!] \neq \emptyset$, take $w \in [\![\varphi]\!]$. Either w is minimal and thus $Min_{\leq}[\![\varphi]\!] \neq \emptyset$ or $M, w \vDash \langle < \rangle \varphi$. Since, by our hypothesis w satisfies Löb Axiom for φ, then $M, w \vDash \langle < \rangle(\varphi \wedge \neg\langle < \rangle \varphi)$, thus there is a world $w' \in W$ s.t. $w' \in [\![\varphi]\!]$ and w' is minimal in this set.

$\Leftarrow$:

Take some world $w \in W$. Suppose $M, w \vDash \langle < \rangle \varphi$, then there is some $w' \in W$, s.t. $w' < w$ and $M, w' \vDash \varphi$. Take $w'' \in Min_{\leq}[\![\varphi]\!] \neq \emptyset$. By minimality, $w'' \leq w'$ and $M, w'' \vDash \varphi \wedge \neg\langle < \rangle \varphi$. By transitivity $w'' < w$, thus $M, w \vDash \langle < \rangle(\varphi \wedge \neg\langle < \rangle \varphi)$. □

Figure 4.1 – Axiomatization $L_{\leq}$ for the Preference Logic $\mathscr{L}_{\leq}(P)$.

$$
\begin{array}{ll}
\mathbf{K}_{\leq}: & [\leq](\varphi \to \psi) \to ([\leq]\varphi \to [\leq]\psi) \\
\mathbf{T}_{\leq}: & [\leq]\varphi \to \varphi \\
\mathbf{4}_{\leq}: & [\leq]\varphi \to [\leq][\leq]\varphi \\
 & \\
\mathbf{K}_{<}: & [<](\varphi \to \psi) \to ([<]\varphi \to [<]\psi) \\
\mathbf{W}_{<}: & [<]([<]\varphi \to \varphi) \to [<]\varphi \\
<\leq_1: & [\leq]\varphi \to [<]\varphi \\
<\leq_2: & [<]\varphi \to [<][\leq]\varphi \\
<\leq_3: & [<]\varphi \to [\leq][<]\varphi \\
<\leq_4: & [\leq]([\leq]\varphi \vee \psi) \wedge [<]\psi \to \varphi \vee [\leq]\psi \\
 & \\
\mathbf{K}_A: & A(\varphi \to \psi) \to (A\varphi \to A\psi) \\
\mathbf{T}_A: & A\varphi \to \varphi \\
\mathbf{4}_A: & A\varphi \to AA\varphi \\
\mathbf{B}_A: & \varphi \to A\neg A\neg\varphi \\
A \leq: & A\varphi \to [\leq]\varphi
\end{array}
$$

Source: the author.

Corollary 4.6 *The logic $L_{\leq}$ depicted in Figure 4.1, taken together with Modus Ponens and the Necessitation Rule for $[\leq]$,$[<]$ and A, is sound and complete for the class of limit preference models, i.e. preference models satisfying the Lewis Limit Assumption.*

We will now dedicate the remainder of this section to prove that this axiomatization $L_{\leq}$ is sound and complete in respect to the class of well-founded preference models.

The problem in providing the completeness result for well-founded preference models for the axiomation of Figure 4.1 is that Löb Axiom (W) is not canonical, i.e. we cannot guarantee that axiomatization is complete by means of the canonical model (BLACKBURN; VAN

BENTHEM; WOLTER, 2006). To do this, we will need to employ the filtrated canonical model, as used to prove completeness of the provability logic GL (or *KW* with the notation employed in our work) (JAPARIDZE; DE JONGH, 1998).

To construct the filtrated canonical model we will need some preliminary definitions. The first one is that of dual of a formula.

Definition 4.7 *Let φ be a formula of $\mathscr{L}_{\leq}$. We define the dual of φ, denoted by the formula $\sim\varphi$ as*

$$\sim\varphi = \begin{cases} \psi & \text{if } \varphi = \neg\psi \text{ for some } \psi \in \mathscr{L}_{\leq} \\ \neg\varphi & \text{otherwise} \end{cases}$$

Another definition we will need is the notion of *extended subformulas* of a formula φ. We need the notion of *extended subformulas* of φ, as opposed to the well-known notion of subformulas, to faithfully represent the entailment relation between the formulas $A\varphi$, $[\leq]\varphi$ and $[<]\varphi$, as required by axioms $A\leq$ and $<\leq_1$ of the axiomatization $L_{\leq}$ in Figure 4.1.

Definition 4.8 *Let φ be a formula of $\mathscr{L}_{\leq}$. We recursively define the extended subformulas of φ as the set:*

$$sub^{+}(\varphi) = \begin{cases} \{\varphi\} & \text{if } \varphi = p \\ \{\varphi\} \cup sub^{+}(\psi) & \text{if } \varphi = \neg\psi \\ \{\varphi\} \cup sub^{+}(\psi) \cup sub^{+}(\xi) & \text{if } \varphi = \psi \wedge \xi \\ \{\varphi, [\leq]\psi, [<]\psi\} \cup sub^{+}(\psi) & \text{if } \varphi = A\psi \\ \{\varphi, A\psi, [<]\psi\} \cup sub^{+}(\psi) & \text{if } \varphi = [\leq]\psi \\ \{\varphi, A\psi, [\leq]\psi\} \cup sub^{+}(\psi) & \text{if } \varphi = [<]\psi \end{cases}$$

With that, we can define how to construct a well-founded preference model that will be used to reason about φ using the axiomatization $L_{\leq}$. This model is similar to the canonical models commonly used in modal correspondence theory (BLACKBURN; VAN BENTHEM; WOLTER, 2006), but limited to finite sets of $L_{\leq}$-theories (or theory bases) as done for the provability logic *KW* (also known as GL, due to Gödel and Löb) (JAPARIDZE; DE JONGH, 1998). These models will be generated by taking parts of the so-called filtrated canonical model of a given formula φ.

Definition 4.9 *Let φ be a satisfiable formula of $\mathscr{L}_{\leq}$. We construct the filtrated canonical model $\mathfrak{M}^{c}_{f}(\varphi) = \langle W, \leq, <, A, v\rangle$ s.t.*

- *The set $U(\varphi) = sub^{+}(\varphi) \cup \{\sim\xi \mid \xi \in sub^{+}(\varphi)\}$ is the universe of φ - a finite set;*

- $W = \{\Delta \subseteq U(\varphi) \mid \Delta$ *is maximally* $L_{\leq} -$ *consistent in* $U(\varphi)\}$ *is the set of maximally consistent parts of* $U(\varphi)$ *that represent the maximal consistent theories of* $\mathscr{L}_{\leq}$*;*
- $\langle \Delta, \Delta' \rangle \in A$ *iff* $\{\varphi \mid A\varphi \in \Delta\} = \{\varphi \mid A\varphi \in \Delta'\}$
- *Let* Δ *be a world, the* $\leq$*-support of* Δ*, denoted by* $s_{\leq}(\Delta)$*, is the set* $s_{\leq}(\Delta) = \{\xi, [\leq]\xi \mid [\leq]\xi \in \Delta\} \cup \{[<]\xi \mid [<]\xi \in \Delta\}$
- $\Delta \leq \Delta'$ *iff* $\langle \Delta, \Delta' \rangle \in A$ *and either* $s_{\leq}(\Delta) = s_{\leq}(\Delta')$ *or both* $s_{\leq}(\Delta') \subseteq s_{\leq}(\Delta)$ *and for all* $[<]\xi \in \Delta'$, $[\leq]\xi \in \Delta$*;*
- $\Delta < \Delta'$ *iff* $\langle \Delta, \Delta' \rangle \in A$ *and for all* $\Box\xi \in \Delta'$, $\xi \in \Delta$ *and* $\Box\xi \in \Delta$*, and also, there is a* $\Box\xi \in \Delta$ *s.t.* $\Box\xi \notin \Delta'$*, with* $\Box \in \{[\leq], [<]\}$*;*
- $v(p) = \{\Delta \in W \mid p \in \Delta\}$

It is clear that $\leq$ is reflexive and transitive from its definition. Also $<$ is irreflexive and $\Delta < \Delta'$ iff $\Delta \leq \Delta'$ and $\Delta' \not\leq \Delta$. More yet, the relation A is an equivalence relation over W. As such, if we take any A-partition $[\Delta]_A = \{\Delta' \mid \langle \Delta, \Delta' \rangle \in A\}$ of the filtrated canonical model $\mathfrak{M}_f^c(\varphi)$, the model $\mathfrak{M}_\Delta = \langle [\Delta]_A, \leq \cap ([\Delta]_A)^2, v \rangle$ defines a well-founded preference model[4].

Now, we only need two auxiliary results that will comprise the central arguments in our completeness proof of Theorem 4.12. The first, Lemma 4.10, states that for any Δ, a world in a filtrated canonical model, $\mathfrak{M}_f^c(\varphi)$ and every $[\leq]\psi \in U(\varphi)$ (c.f. Definition 4.9), if $\mathfrak{M}_\Delta, \Delta \vDash [\leq]\psi$, then $[\leq]\psi \in \Delta$. This is done by showing that if $\neg[\leq]\psi \in \Delta$ then there must be a world $\Delta' \in [\Delta]_A$ s.t. $\sim \psi \in \Delta'$ and $\Delta' \leq \Delta$.

Lemma 4.10 *Let* $\mathfrak{M}_f^c(\varphi) = \langle W, \leq, <, A, v \rangle$ *be a filtrated canonical model as above and* $\Delta \in W$ *s.t.* $\neg[\leq]\psi \in \Delta$*. There is* $\Delta' \in W$ *s.t.* $\sim \psi \in \Delta'$ *and* $\Delta' \leq \Delta$*.*

Proof: We have two cases consider, the case in which $[<]\psi \notin \Delta$ and the case $[<]\psi \in \Delta$.
First case: $([<]\psi \notin \Delta)$

In the first case, $\neg[<]\psi \in \Delta$ by maximality of Δ.

Suppose the set $s_{\leq}(\Delta) \cup \{[\leq]\beta \mid [<]\beta \in \Delta\} \cup \{A\beta \mid A\beta \in \Delta\} \cup \{\neg A\beta \mid \neg A\beta \in \Delta\} \cup \{\sim \psi\}$ is inconsistent. As such there are a finite sets $\{[\leq]\alpha_1, \ldots, [\leq]\alpha_P\}$, with $[\leq]\alpha_i \in \Delta$ or $[<]\alpha_i \in \Delta$ for each i,

[4]Well-foundedness comes from the fact that $[\Delta]_A$ is finite and $<$ is irreflexive.

and $\{A\beta_1,\ldots,A\beta_p\} \subseteq \Delta$ and $\{\neg A\theta_1,\ldots,\neg A\theta_q\} \subseteq \Delta$ s.t.

1	$\bigwedge_1^p A\beta_i \wedge \bigwedge_1^q \neg A\theta_i \wedge \bigwedge_1^n [\leq]\alpha_i \wedge \sim \psi \rightarrow \bot$	by definition of inconsistency
2	$\bigwedge_1^p A\beta_i \wedge \bigwedge_1^q \neg A\theta_i \wedge \bigwedge_1^n [\leq]\alpha_i \rightarrow \psi$	*reductio ad absurdum*
3	$[<](\bigwedge_1^p A\beta_i \wedge \bigwedge_1^q \neg A\theta_i \wedge \bigwedge_1^n [\leq]\alpha_i) \rightarrow [<]\psi$	by $(\varphi \rightarrow \psi) \rightarrow ([<]\varphi \rightarrow [<]\psi)$
4	$\bigwedge_1^p [<]A\beta_i \wedge \bigwedge_1^q [<]\neg A\theta_i \wedge \bigwedge_1^n [<][\leq]\alpha_i \rightarrow [<]\psi$	by $[\leq](\varphi \wedge \psi) \leftrightarrow [\leq]\varphi \wedge [\leq]\psi$
5	$\bigwedge_1^p [<]A\beta_i \wedge \bigwedge_1^q [<]\neg A\theta_i \wedge \bigwedge_1^n [<]\alpha_i \rightarrow [<]\psi$	by $[<]\varphi \rightarrow [<][\leq]\varphi$
6	$\bigwedge_1^p A\beta_i \wedge \bigwedge_1^q [<]\neg A\theta_i \wedge \bigwedge_1^n [<]\alpha_i \rightarrow [<]\psi$	by $A\varphi \rightarrow [<]\varphi$ and $A\varphi \rightarrow AA\varphi$
7	$\bigwedge_1^p A\beta_i \wedge \bigwedge_1^q \neg A\theta_i \wedge \bigwedge_1^n [<]\alpha_i \rightarrow [<]\psi$	by $A\varphi \rightarrow [<]\varphi$ and $\neg A\varphi \rightarrow A\neg A\varphi$

Well, by construction, $[<]\alpha_i \in \Delta$, since $[\leq]\varphi \rightarrow [<]\varphi$, $A\beta_i \in \Delta$ and $\neg A\theta_i \in \Delta$, for all i, thus $[<]\psi \in \Delta$, which is a contradiction to our assumption. Thus $\delta' = s_{\leq}(\Delta) \cup \{[\leq]\beta \mid [<]\beta \in \Delta\} \cup \{A\beta \mid A\beta \in \Delta\} \cup \{\neg A\beta \mid \neg A\beta \in \Delta\} \cup \{\sim \psi\}$ must be consistent and, as such, there is a consistent extension Δ' of δ'. Since $\{A\beta \mid A\beta \in \Delta\} \cup \{\neg A\beta \mid \neg A\beta \in \Delta\} \subseteq \Delta'$, by definition of A, $\langle \Delta, \Delta' \rangle \in A$. Thus, by definition of $\leq$, $\Delta' \leq \Delta$.

Second case: $([<]\psi \in \Delta)$

This is the tricky case. As $[<]\psi \in \Delta$ then, if there is a $\Delta' \leq \Delta$ with $\sim \psi \in \Delta'$, then it must be the case that $\Delta \leq \Delta'$. We have to show then that the set

$$\delta' = s_{\leq}(\Delta) \cup \{\neg[\leq]\xi \mid \neg[<]\xi \in \Delta\} \cup \{\neg[<]\xi \mid \neg[<]\xi \in \Delta\} \cup \{A\varphi \mid A\varphi \in \Delta\} \cup \{\neg A\beta \mid \neg A\beta \in \Delta\} \cup \{\sim \psi\}$$

is $L_{\leq}$-consistent.

Imagine δ' is inconsistent. Again, there must be finite sets $\{[\leq]\alpha_1,\ldots,[\leq]\alpha_P\}$, $\{[\leq]\beta_1,\ldots,[\leq]\beta_m\}$, $\{\neg[\leq]\gamma_1,\ldots,\neg[\leq]\gamma_p\}$, $\{\neg[\leq]\xi_1,\ldots,\neg[\leq]\xi_q\}$, $\{A\zeta_1,\ldots,A\zeta_r\}$ and $\{\neg A\rho_1,\ldots,A\rho_s\}$ s.t.

1	$\left(\begin{array}{l}\bigwedge_1^r A\zeta_i \wedge \bigwedge_1^s \neg A\rho_i \wedge \bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m [\leq]\beta_i \wedge \\ \bigwedge_1^p \neg[\leq]\gamma_i \wedge \bigwedge_1^q \neg[<]\xi_i \wedge \sim\psi\end{array}\right) \to \bot$	by definition of inconsistency
2	$\left(\begin{array}{l}\bigwedge_1^r A\zeta_i \wedge \bigwedge_1^s \neg A\rho_i \wedge \bigwedge_1^n [\leq]\alpha_i \wedge \\ \bigwedge_1^m [\leq]\beta_i \wedge \bigwedge_1^p \neg[\leq]\gamma_i \wedge \bigwedge_1^q \neg[<]\xi_i\end{array}\right) \to \psi$	by *reductio ad absurdum*
3	$\left(\begin{array}{l}\bigwedge_1^r A\zeta_i \wedge \bigwedge_1^s \neg A\rho_i \wedge \\ \bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m [\leq]\beta_i\end{array}\right) \to ((\bigwedge_1^p \neg[\leq]\gamma_i \wedge \bigwedge_1^q \neg[<]\xi_i) \to \psi)$	by $(\varphi \wedge \psi \to \xi) \leftrightarrow (\varphi \to (\psi \to \xi))$
4	$\left(\begin{array}{l}\bigwedge_1^r A\zeta_i \wedge \bigwedge_1^s \neg A\rho_i \wedge \\ \bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m [\leq]\beta_i\end{array}\right) \to (\bigvee_1^p [\leq]\gamma_i \vee \bigvee_1^q [<]\xi_i \vee \psi)$	by $(\varphi \to \psi) \leftrightarrow (\neg\varphi \vee \psi)$
5	$\left(\begin{array}{l}[<](\bigwedge_1^r A\zeta_i \wedge \bigwedge_1^s \neg A\rho_i \wedge \\ \bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m [\leq]\beta_i)\end{array}\right) \to [\leq](\bigvee_1^p [\leq]\gamma_i \vee \bigvee_1^q [<]\xi_i \vee \psi)$	by $(\varphi \to \psi) \to ([<]\varphi \to [<]\psi)$
6	$\left(\begin{array}{l}\bigwedge_1^r [\leq]A\zeta_i \wedge \bigwedge_1^s [\leq]\neg A\rho_i \wedge \\ \bigwedge_1^n [<][\leq]\alpha_i \wedge \bigwedge_1^m [<][<]\beta_i\end{array}\right) \to [\leq](\bigvee_1^p [\leq]\gamma_i \vee \bigvee_1^q [<]\xi_i \vee \psi)$	by $[\leq](\varphi \wedge \psi) \leftrightarrow [\leq]\varphi \wedge [\leq]\psi$
7	$\left(\begin{array}{l}\bigwedge_1^r [\leq]A\zeta_i \wedge \bigwedge_1^s [\leq]\neg A\rho_i \wedge \\ \bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m [<]\beta_i\end{array}\right) \to [\leq](\bigvee_1^p [\leq]\gamma_i \vee \bigvee_1^q [<]\xi_i \vee \psi)$	by $[\leq]\varphi \to [<][\leq]\varphi$ and $[<]\varphi \to [<][<]\varphi$
8	$\left(\begin{array}{l}\bigwedge_1^r A\zeta_i \wedge \bigwedge_1^s [\leq]\neg A\rho_i \wedge \\ \bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m [<]\beta_i\end{array}\right) \to [\leq](\bigvee_1^p [\leq]\gamma_i \vee \bigvee_1^q [<]\xi_i \vee \psi)$	by $A\varphi \to [\leq]\varphi$ and $A\varphi \to AA\varphi$
9	$\left(\begin{array}{l}\bigwedge_1^r A\zeta_i \wedge \bigwedge_1^s \neg A\rho_i \wedge \\ \bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m [<]\beta_i\end{array}\right) \to [\leq](\bigvee_1^p [\leq]\gamma_i \vee \bigvee_1^q [<]\xi_i \vee \psi)$	by $A\varphi \to [\leq]\varphi$ and $\neg A\varphi \to A\neg A\varphi$
10	$\Lambda \to [\leq](\bigvee_1^p [\leq]\gamma_i \vee \bigvee_1^q [<]\xi_i \vee \psi)$	let $\Lambda = \bigwedge_1^r A\zeta_i \wedge \bigwedge_1^s \neg A\rho_i \wedge \bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m [<]\beta_i$
11	$\Lambda \to [\leq](\bigvee_1^q [<]\xi_i \vee \bigvee_1^p [\leq]\gamma_i \vee \psi)$	by $(A \vee B) \leftrightarrow (B \vee A)$
12	$\Lambda \to [\leq](\bigvee_1^q [\leq][<]\xi_i \vee \bigvee_1^p [\leq]\gamma_i \vee \psi)$	by $[<]A \to [\leq][<]A$
13	$\Lambda \to [\leq]([\leq](\bigvee_1^q [<]\xi_i) \vee \bigvee_1^p [\leq]\gamma_i \vee \psi)$	by $[<]A \vee [<]B \to [<](A \vee B)$
14	$\Lambda \wedge [<]\psi \to [\leq]([\leq](\bigvee_1^q [<]\xi_i) \vee \bigvee_1^p [\leq]\gamma_i \vee \psi) \wedge [<]\psi$	by $(A \to B) \to (A \wedge C \to B \wedge C)$
15	$\Lambda \wedge [<]\psi \to [\leq]([\leq](\bigvee_1^q [<]\xi_i) \vee \bigvee_1^p [\leq]\gamma_i \vee \psi) \wedge [<](\bigvee_1^p [\leq]\gamma_i \vee \psi)$	by $[\leq]A \to [\leq](A \vee B)$ by $<\leq_4$ in Figure 4.1
16	$\Lambda \wedge [<]\psi \to (\bigvee_1^q [<]\xi_i) \vee [\leq](\bigvee_1^p [\leq]\gamma_i \vee \psi)$	substituting φ for $\bigvee_1^q [<]\xi_i$ and ψ for $\bigvee_1^p [\leq]\gamma_i \vee \psi$

As before, $\Delta \vdash \Lambda$, since all $[\leq]\alpha_i \in \Delta$, $[<]\beta_i \in \Delta$,$A\zeta_i \in \Delta$ and $[<]\rho_i \in \Delta$. Also, by assumption,

$[<]\psi \in \Delta$. Thus,

$$\Delta \vdash (\bigvee_1^q [<]\xi_i) \vee [\leq](\bigvee_1^p [\leq]\gamma_i \vee \psi) \qquad (*)$$

.

But, by construction, $\Delta \vdash \bigwedge_1^q \neg[<]\xi_i$ and $\Delta \vdash \bigwedge_1^p \neg[\leq]\gamma_i$, then $\Delta \vdash \neg\bigvee_1^q [<]\xi_i$ and $\Delta \vdash \neg\bigvee_1^p [\leq]\gamma_i$. Also, by hypothesis $\Delta \vdash \neg[\leq]\psi$. As such,

1	$\Delta \vdash \neg\bigvee_1^p [\leq]\gamma_i \wedge \neg[\leq]\psi$	by the facts that $\Delta \vdash \neg\bigvee_1^p [\leq]\gamma_i$ and $\Delta \vdash \neg[\leq]\psi$
2	$\Delta \vdash \neg(\bigvee_1^p [\leq]\gamma_i \vee [\leq]\psi)$	$\neg A \wedge \neg B \leftrightarrow \neg(A \vee B)$
3	$\Delta \vdash \neg[\leq](\bigvee_1^p \gamma_i \vee \psi)$	$[\leq]A \vee [\leq]B \rightarrow [\leq](A \vee B)$
4	$\Delta \vdash \neg(\bigvee_1^q [<]\xi_i) \wedge \neg[\leq](\bigvee_1^p \gamma_i \vee \psi)$	by 3 and the fact that $\Delta \vdash \neg\bigvee_1^q [<]\xi_i$
5	$\Delta \vdash \neg((\bigvee_1^q [<]\xi_i) \vee [\leq](\bigvee_1^p \gamma_i \vee \psi))$	by $\neg A \wedge \neg B \rightarrow \neg(A \vee B)$

But this is a contradiction to $(*)$ above, since Δ is $L_{\leq}$-consistent by definition. Thus, we must conclude that the set δ' is $L_{\leq}$-consistent. As such, there is a maximally consistent extension Δ' of δ' and since $\delta' \subseteq \Delta'$, we have that $\langle \Delta, \Delta' \rangle \in A$, $\sim\psi \in \Delta'$ and $s_{\leq}(\Delta') = s_{\leq}(\Delta)$. As such, $\Delta' \leq \Delta$. □

The second result we need, Lemma 4.11, is similar to the previous one, only we show that for any world Δ in a filtrated canonical model $\mathfrak{M}_f^c(\varphi)$ and every $[<]\psi \in U(\varphi)$, if $\mathfrak{M}_\Delta, \Delta \vDash [<]\psi$ then $[<]\psi \in \Delta$.

Lemma 4.11 *Let $\mathfrak{M}_f^c = \langle W, \leq, <, A, v\rangle$ be a filtrated canonical model as above and $\Delta \in W$ s.t. $\neg[<]\psi \in \Delta$, then the set $\delta' = \{\alpha, [\leq]\alpha \mid [\leq]\alpha \in \Delta$ or $[<]\alpha \in \Delta\} \cup \{A\alpha \mid A\alpha \in \Delta\} \cup \{A\alpha \mid A\alpha \in \Delta\} \cup \{\sim\psi, [<]\psi\}$ is $L_{\leq}$-consistent.*

Proof:

Suppose not, then there is are finite subsets $\{[\leq]\alpha_1, \ldots, [\leq]\alpha_P\}$, $\{A\beta_1, \ldots, A\beta_m\}$ and $\{\neg A\rho_1, \ldots, \neg A\rho_k\}$ s.t.:

1	$\bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m A\beta_i \wedge \bigwedge_1^k \neg A\rho_i \wedge [<]\psi \wedge \sim\psi \rightarrow \bot$	by definition of inconsistency
2	$\bigwedge_1^n [\leq]\alpha_i \wedge [<]\psi \rightarrow \psi$	*reduction ad absurdum*
3	$\bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m A\beta_i \wedge \bigwedge_1^k \neg A\rho_i \rightarrow ([<]\psi \rightarrow \psi)$	by $(\varphi \wedge \theta \rightarrow \xi) \leftrightarrow (\varphi \rightarrow (\theta \rightarrow \xi))$
4	$[<](\bigwedge_1^n [\leq]\alpha_i \wedge \bigwedge_1^m A\beta_i \wedge \bigwedge_1^k \neg A\rho_i) \rightarrow [<]([<]\psi \rightarrow \psi)$	by $(\varphi \rightarrow \xi) \rightarrow ([<]\varphi \rightarrow [<]\xi)$
5	$\bigwedge_1^n [<][\leq]\alpha_i \wedge \bigwedge_1^m [<]A\beta_i \wedge \bigwedge_1^k [<]\neg A\rho_i \rightarrow [<]\psi$	by W in Figure 4.1
6	$\bigwedge_1^n [<]\alpha_i \wedge \bigwedge_1^m [<]A\beta_i \wedge \bigwedge_1^k [<]\neg A\rho_i \rightarrow [<]\psi$	by $[<]\varphi \rightarrow [<][\leq]\varphi$
7	$\bigwedge_1^n [<]\alpha_i \wedge \bigwedge_1^m A\beta_i \wedge \bigwedge_1^k [<]\neg A\rho_i \rightarrow [<]\psi$	by $A\varphi \rightarrow [<]\varphi$ and $A\varphi \rightarrow AA\varphi$
8	$\bigwedge_1^n [<]\alpha_i \wedge \bigwedge_1^m A\beta_i \wedge \bigwedge_1^k \neg A\rho_i \rightarrow [<]\psi$	by $A\varphi \rightarrow [<]\varphi$ and $\neg A\varphi \rightarrow A\neg A\varphi$

But all $A\beta_i$, $\neg A\rho_i$ and $[<]\alpha_i$ are in Δ, since $[\leq]\varphi \rightarrow [<]\varphi$, thus by maximality $[<]\psi \in \Delta$, which is

a contradiction to our initial assumption. We must then conclude that δ' is $L_{\leq}$-consistent. As such, there is a maximal consistent extension Δ' of δ', s.t. $\langle \Delta, \Delta' \rangle \in A$, $\sim \psi \in \Delta'$, $\{\alpha, [\leq]\alpha \mid [\leq]\alpha \in \Delta \text{ or } [<]\alpha \in \Delta\} \subseteq \Delta'$ and there is $[<]\psi \in \Delta'$ s.t. $[<]\psi \notin \Delta$. In other words, by the definition of $<$, $\Delta' < \Delta$. □

With this two fundamental results, we can finally prove completeness of the axiomatization $L_{\leq}$ proposed in Figure 4.1 with respect to the class of preference models with a well-founded strict part $<$. In the proof below note that soundness follows easily from Corollary 4.6.

Theorem 4.12 *The axiomatization $L_{\leq}$ proposed in Figure 4.1 with all propositional tautologies, modus ponens and necessitation rules is sound and complete with respect to preference models with a well-founded strict part $<$.*

Proof: The soundness of the axioms is easy. Since all well-founded model satisfy the Lewis Limit Assumption, by Corollary 4.6, $L_{\leq}$ is sound wrt well-founded preference models. Now, we must only concern ourselves with completeness.

We will show this by the contrapositive. Let φ be a formula s.t. $L_{\leq} \nvdash \varphi$, i.e. φ is not a theorem of $L_{\leq}$. Take the filtrated canonical model $\mathfrak{M}^c_f(\varphi) = \langle W, \leq, <, A, v \rangle$.

Let's prove that for any $\xi \in U(\varphi)$, $\xi \in \Delta$ iff $\mathfrak{M}_\Delta, \Delta \vDash \xi$. This is done by induction on the structure of ξ.

The interesting cases are $\xi = [\leq]\psi$ and $\xi = [<]\psi$. Let's first show the case $\xi = [\leq]\psi$.

($\Rightarrow$):
Assume $[\leq]\psi \in \Delta$, then by definition of $\leq$, for any $\Delta' \leq \Delta$, $\psi \in \Delta'$. So, $\mathfrak{M}^c_f, \Delta \vDash [\leq]\psi$

($\Leftarrow$):
To prove this case, we will use the contrapositive. Assume $\neg[\leq]\psi \in \Delta$, then by Lemma 4.10, there is a $\Delta' \leq \Delta$ such that $\psi \notin \Delta'$, thus $\mathfrak{M}^c_f, \Delta \nvDash [\leq]\psi$.

We now must examine the case of $\xi = [<]\psi$.

($\Rightarrow$):
Assume $[<]\psi \in \Delta$, then by definition of $<$, for any $\Delta' < \Delta$, $\psi \in \Delta'$. So, $\mathfrak{M}^c_f, \Delta \vDash [<]\psi$

($\Leftarrow$):
To prove this case, we will use the contrapositive. Assume $\neg[<]\psi \in \Delta$, then by Lemma 4.11, there is a $\Delta' < \Delta$ such that $\psi \notin \Delta'$, thus $\mathfrak{M}^c_f, \Delta \nvDash [<]\psi$.

In particular, $\varphi \in U(\varphi)$ and, since $L_{\leq} \nvdash \varphi$ there is a $\Delta \in W$, s.t. $\sim \varphi \in \Delta$ as W contains all maximally $L_{\leq}$-consistent subsets of $U(\varphi)$. Thus, $\mathfrak{M}^c_f, \Delta \nvDash \varphi$, i.e. φ is not valid in regards to well-founded preference models. □

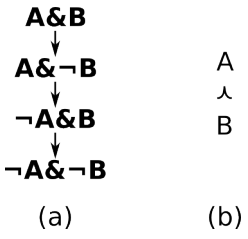
4.3 Preferences and priority graphs

Having paved the way for the dynamification of preference logic providing a sound and complete axiomatization for (static) preference logic restricted to well-founded models, in this section we will present the concept of syntactic structures know as priority graphs. These structures will be used as computational-friendly representation of preference models used in the remainder of the work.

Representations of preference-like relations, such as ideality and plausibility, as ordering relations between formulas have an extensive history in the formal treatment of mental attitudes. For deontic systems, Van Benthem, Grossi and Liu (2014) provides a resumed history of its use. For Belief Change, such representations occurs, for example, as entrenchment relations (GÄRDENFORS; MAKINSON, 1988) and as (stratified) belief bases (ROTT, 98; ROTT, 2009) which have been widely used to represent an agent's epistemic state. They can also be be seen in representation of desires (DOYLE; SHOHAM; WELLMAN, 1991; LANG; VAN DER TORRE; WEYDERT, 2003) and in non-monotonic reasoning (BENFERHAT et al., 1993; KACI; VAN DER TORRE, 2005)

Liu (2011) introduces the syntactic-based structures of priority graphs (or shortly P-graphs). A P-graph is, in essence, a partial order over propositional sentences, which is used to represent some agent's preferences in regards to a subject. This simple idea, which lies in the core of several previous syntactic representations of preferences in the literature, allows a compact way of representing a preference relation, see for example Figure 4.2 where the preference order of the model depicted in (a) is encoded in the priority-graph in (b)[5]. Notice in the Figure 4.2, an edge starting in s_1 and ending in s_2 means $s_1 \leq s_2$. Also, A and B are the propositional symbols of the set P and a world in which A and B are satisfied is represented as $A\&B$.

Figure 4.2 – A preference model (a) and an equivalent P-graph (b)



Source: the author

[5] How to achieve one representation from the other will be discussed in a while.

We believe priority graphs to be easily embedded in computational systems. As such, by the connection between Preference Logic and priority graphs, we can also define computational methods to reason about preferences (and later preference dynamics) based on them. In fact, in Chapter 6, we will use priority graphs as representation structures to encode an agent program state and to reason about the agent's mental state.

Let's begin our discussion by introducing the notion of priority graph (or P-graph as we will call them).

Definition 4.13 *(LIU, 2011) Let $\mathscr{L}_0(P)$ be the propositional language constructed over the set of propositional letters P, as usual. A P-graph is a tuple $\mathscr{G} = \langle \Phi, \prec \rangle$ where $\Phi \subset \mathscr{L}_0(P)$, is a set of propositional sentences and $\prec$ is a strict partial order on Φ.*

As we do for preference models, $\varphi \prec \psi$ will be understood as '*the agent prefers that φ than ψ*'. Similarly to the order relation in preference models, we can give several interpretations for priority graphs, depending on the phenomena being encoded. For example, if we wish to encode agents desires by means of a P-graph, the information $\varphi \prec \psi$ may be interpreted as φ *is more desirable (or wanted) than* ψ and can be viewed as a prioritization over the agent's desires. If we wish to represent deontic phenomena, however, the same information may be interpreted as φ *is (deontically) better than* ψ.

Priority graphs define an order relation among sentences whilst preference models define an order relation among possible worlds. The question remaining is how those two orderings can be connected. Given a valuation function of propositions over a set of possible worlds, an ordering among sentences as defined by a P-graph can be lifted to an order over worlds. In fact, there are many ways of establishing such an ordering over worlds (LANG; VAN DER TORRE; WEYDERT, 2003).

If we consider priority graphs as justification structures, such as used in Truth Maintenance systems (DOYLE, 1979), the lexicographic ordering[6] of worlds according the priority graph seems like a ideal candidate for such a lifting. In other words, if we take the information $\varphi \prec \psi$ as a reason (or justification) for believing that any world satisfying φ to be preferable to any world satisfying $\neg\varphi \wedge \psi$, then we can lift the relation $\prec$ to a relation among worlds by using the lexicographic ordering. For example, take the model in Figure 4.2 (a), the order in this model has been lifted from the order in the P-graph depicted in Figure 4.2 (b). Since $A \prec B$ in the P-graph, any A-world (i.e. $A\&B$ and $A\&\neg B$) is preferred to the world satisfying B but

[6]By lexicographic ordering we mean that the order defined over the worlds is relative to the order of which formulas in the P-graph they satisfy, i.e. if $\varphi < \psi$ in the P-graph, then any world w satisfying φ must be preferred to any world w' not satisfying φ but satisfying ψ.

not A ($\neg A\&B$). Also, $A\&B$ is preferred to $A\&\neg B$ since the former satisfies both nodes A and B in the priority graph, while the later only satisfies the node A. The world $\neg A\&\neg B$ is the least preferred world since it does not satisfy any of the nodes in the priority graph.

Definition 4.14 *(LIU, 2011) Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-graph, W be a finite non-empty set of states or possible worlds, and $v : P \to 2^W$ be a valuation function. The order relation $\leq_{\mathscr{G}} \subseteq W^2$ is defined as follows:*

$$w \leq_{\mathscr{G}} w' \quad \textit{iff} \quad \forall \varphi \in \Phi : (w' \vDash \varphi \Rightarrow w \vDash \varphi) \vee (\exists \psi \prec \varphi : (w \vDash \psi \textit{ and } w' \nvDash \psi))$$

Notice that the condition above corresponds to the lexicographic ordering based on the graph $\mathscr{G}$, since $\langle w, w' \rangle \in \leq_{\mathscr{G}}$ iff either both w and w' satisfy exactly the same formulas of the graph, or, if there is a φ that w' satisfies and w does not, then there is a formula ψ preferable to φ, s.t. w satisfies it and w' does not.

Van Benthem, Grossi and Liu (2014) has shown that such an order, defined as above, has some very desirable properties, such as Fact 4.15 and Theorem 4.17 below, that will allow the connection of priority graphs to our previously studied preference models.

Fact 4.15 *(VAN BENTHEM; GROSSI; LIU, 2014) The relation $\leq_{\mathscr{G}}$, as defined above, is a preference relation whose strict part is well-founded.*

From Definition 4.14 and Fact 4.15, we can see that a priority graph induces a well-founded preference relation over a set of possible worlds. As such, we will say that a model $M = \langle W, \leq_{\mathscr{G}}, v \rangle$ is induced by a given priority graph $\mathscr{G}$, if its preference relation is constructed by lifting the relation of the graph $\mathscr{G}$, according to the Definition 4.14.

Definition 4.16 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ a P-graph and $M = \langle W, \leq, v \rangle$ a preference model. We say M is induced by $\mathscr{G}$ iff $\leq = \leq_{\mathscr{G}}$, where $\leq_{\mathscr{G}}$ is the relation defined in Definition 4.14 over the set W considering the valuation v.*

Liu (2011) shows that any model with a reflexive, transitive relation is induced by some priority graph.

Theorem 4.17 *(LIU, 2011) Let $M = \langle W, R, v \rangle$ a modal model. The following two statements are equivalent:*

1. *$M = \langle W, R, v \rangle$ is a preference model[7];*

[7] Liu (2011) in her presentation does not requires well-foundedness of the models. Later, Van Benthem, Grossi and Liu (2014) prove that if the set Φ is finite, then the model will be well-founded. In our presentation we will maintain well-foundedness, without prejudice to Liu's original proof.

2. *There is a priority graph* $\mathscr{G} = (\Phi, \prec)$ *s.t.* $\forall w, w' \in W :$ wRw' *iff* $w \leq_{\mathscr{G}} w'$.

More yet, if W *is finite, then so is* Φ.

The Theorem 4.17 shows, thus, that P-graphs and preference models are two ways of encoding preferences. While the construction of a preference relation from a P-graph was established, the construction of a P-graph from a preference model has not been discussed. This construction is carried by taking all the worlds in a same equivalence cluster in the preference relation $\leq$ and representing by a propositional formula satisfied only by those worlds. For most cases, this formula can be representing each world as a propositional formula and taking their disjunction. For some models, however, it is necessary to extend the propositional symbol set P and construct the graph to the model in this extended logic that is isomorphic to the original one.

4.3.1 Representing conditional preferences by means of P-graphs

Having introduced the notion of priority graphs, our objective is to be able to use these structures as means to reason about preferences. This way, we can use priority graphs as data structures in computational systems - particularly in the implementation of an agent programming language. As such, we begin the investigation of reasoning with priority graphs, focusing on an important subset of formulas of our language $\mathscr{L}_{\leq}(P)$, namely conditional preferences, introduced in Definition 4.4.

Let's remember, a conditional preference that ψ given φ, i.e.'*in the most preferred* φ*-worlds,* ψ *holds*,' is expressed by the formula $C(\psi|\varphi)$ defined in Preference Logic as the formula $A(\mu\varphi \rightarrow \psi)$.

As we discussed in Section 4.1, conditional preferences are often regarded as faithful encodings of (some) mental attitudes. In fact, we shall use these formulas, in Chapter 5, to encode the mental attitudes of the BDI architecture, which will base our logical investigation of Agent Programming. Luckily, Van Benthem, Grossi and Liu (2014) show that one can encode conditional preferences using P-graphs.

To construct a graph-based codification for such a formula, Van Benthem, Grossi and Liu (2014) use the notion of maximal paths in a graph.

Definition 4.18 *Let* $\mathscr{G} = \langle \Phi, \prec \rangle$ *be a finite P-graph, i.e.* Φ *is finite. We say the sequence* $\langle \varphi_1, \ldots, \varphi_P \rangle$*, with* $\varphi_i \in \Phi$*, is a maximal path in* $\mathscr{G}$*, denoted by* $\langle \varphi_1, \ldots, \varphi_P \rangle \in \Pi_{\mathscr{G}}$ *iff:*

- φ_1 *is minimal in* $\mathscr{G}$*, i.e. there is no* $\varphi \in \Phi$ *s.t.* $\varphi \prec \varphi_1$
- φ_P *is final in* $\mathscr{G}$*, i.e. there is no* $\varphi \in \Phi$ *s.t.* $\varphi_P \prec \varphi$
- $\langle \varphi_1, \ldots, \varphi_P \rangle$ *is a chain, i.e. for all* $1 \leq i < n$ $\varphi_i \prec \varphi_{i+1}$ *and there is no* $\varphi \in \Phi$ *s.t.* $\varphi_i \prec \varphi \prec \varphi_{i+1}$.

The maximal paths of a P-graph describe the total suborders of an induced model, i.e. all the total orders contained in the accessibility relation of the induced model. Well, the minimal worlds of the induced model are exactly the union of the minimal worlds in each total suborder.

Using the linear order of a maximal path, Van Benthem, Grossi and Liu (2014) provide the following encoding for the minimal worlds satisfying φ in a maximal path $\langle \varphi_1, \ldots, \varphi_P \rangle$, denoted by the formula 4.1.

$$\mu_{\langle}\varphi_1, \ldots, \varphi_P\rangle(\varphi) \equiv \bigwedge_{1 \leq i \leq n} ((E((\bigvee_{1 \leq j \leq i} \varphi_j) \wedge \varphi)) \rightarrow ((\bigvee_{1 \leq j \leq i} \varphi_j) \wedge \varphi)) \tag{4.1}$$

With that formula, to encode the notion defined in the conditional preference $C(\psi|\varphi)$ as given in the Definition 4.4, we only need to take the best worlds satisfying φ in all the paths in the graph and guarantee they satisfy ψ. We will then define the P-graph based conditional preference $C_{\mathscr{G}}(\psi|\varphi)$

$$C_{\mathscr{G}}(\psi|\varphi) \equiv A((\bigvee_{\langle \varphi_1, \ldots, \varphi_P \rangle \in \Pi_{\mathscr{G}}} \mu_{\langle \varphi_1, \ldots, \varphi_P \rangle} \varphi) \rightarrow \psi) \tag{4.2}$$

With that, Van Benthem, Grossi and Liu (2014) show that the graph-based conditional preference of formula 4.2 and the conditional preference defined in Definition 4.4 are equivalent.

Proposition 4.19 *Let* $\mathscr{G}$ *be a P-graph,* $M = \langle W, \leq, v \rangle$ *a preference model induced by* $\mathscr{G}$ *and* $w \in W$ *a world in* M.

$$M, w \vDash C(\psi|\varphi) \quad \textit{iff} \quad M, w \vDash C_{\mathscr{G}}(\psi|\varphi)$$

As we said previously, in Chapter 5 we will use conditional preferences to encode mental attitudes used in Agent Programming. We propose, in Chapter 6, that priority graphs can be used as a data structure to implement the semantics of the language, as such we need a way to reason about conditional statements such as $C(\psi \mid \varphi)$ using priority graphs, i.e. we can infer the satisfiability of a formula $C(\psi \mid \varphi)$ without reference to a preference model.

Proposition 4.19 above is a initial step in this direction, but notice that the formula 4.1 uses existential modalities E whose satisfiability is intrinsically dependent on the model M. To solve this 'problem' we will introduce some interesting models, which we call broad models.

Definition 4.20 *We say a preference model $M = \langle W, \leq, v \rangle$ is broad if the function $f : W \to 2^P$ s.t. $f(w) = \{p \in P \mid w \in v(p)\}$ is bijective.*

Broad models are those in which all possible truth assignments for the propositional symbols of P are represented. In other words, for any valuation for the propositional symbols in P, there is a world $w \in W$ satisfying exactly the same propositional symbols of P to which this valuation attributes a true value.

These models present some interesting properties for our study. The first of these properites is that in a broad model the universal and existential modalities A and E, respectively, represent propositional validity and satisfiability.

Fact 4.21 *Let $M = \langle W, \leq, v \rangle$ be a broad model and $\varphi \in \mathscr{L}_{\leq}$ without order modalities, i.e. neither $[\leq]$ nor $[<]$ occur in φ. For any $w \in W$, it holds that*

$$M, w \vDash E\varphi \quad \textit{iff} \quad \textit{there is a propositional valuation } v' s.t. v'(\varphi) = 1$$
$$M, w \vDash A\varphi \quad \textit{iff} \quad \textit{for all propositional valuations } v' \textit{ it holds that } v'(\varphi) = 1$$

As such, since order modalities do not appear in the formula 4.1, if we only consider broad models, by Fact 4.21, conditional preferences based on P-graphs - as presented in Definition 4.2 - can be decided using boolean satifiability.

This means that, in Chapter 6, if we take broad models as the desired logical representation of the agent's mental state, we can use priority graphs as means to implement an agent programming language. Other interesting properties of broad models are investigated on Section 4.6, where we use broad models to study contraction operations.

In the next Section, we present some update operations already discussed in the literature. Then in Section 4.5, we extend the logic with contraction operators based on the study of iterated belief change (RAMACHANDRAN; NAYAK; ORGUN, 2012). This extension of Preference Logic with contraction operators is, as far as we know, a novel contribution of this thesis. While working on this extension with contraction operators we were also able to proof a negative result about the limits of harmony between semantic models and syntactic priority graphs.

4.4 Dynamifying preference logic: update operators

In this section, we present three well-studied operations representing different update mechanisms in agents preferences. Namely, we study the operations of public announcement,

radical revision and public suggestion. In the following discussion, for each operation we introduce a modality representing it. For the case of public announcement, for instance, we introduce formulas of the type $[!\varphi]\psi$, where φ is a propositional formula and ψ is any formula of this extended language. Notice that we require that φ in the argument of announcement $!\varphi$ is a propositional formula. This is not necessary for most of the results presented below, but this choice guarantees that the logic resulting of the addition of public announcements to be definable by means of operations on priority graphs (c.f.Theorem 4.27). The same expressive restriction will be employed in every dynamic operator we will present in this chapter.

Throughout both this section and Section 4.5, we will follow the methodology sketched by Van Benthem and Liu (2007) for the study of dynamic operations in Dynamic Epistemic Logic. Namely, when studying a given dynamic operator, we will introduce this operator by means of a transformation on preference models and give a representation for it by means of a PDL program. From the PDL representation, we will derive a set of reduction axioms for the operation by means of the technique described in Fact 4.22 below. Finally, we will investigate the possibility of defining the operations by means of syntactic transformations of P-graphs providing the transformation when possible.

The result below describes the technique that will be used in the following sections to derive reduction axioms for the operations we will study in this chapter, i.e. axioms interpret the formulas of the extended language within the static language of preferences. These reduction axioms will be used to provide a complete axiomatization of the Dynamic Preference Logic we propose in this chapter.

Fact 4.22 *(VAN BENTHEM; LIU, 2007) Every relation-changing operation that is definable in PDL without iteration has a complete set of reduction axioms in Dynamic Preference Logic.*

Proof: Let σ be an operation changing the basic accessibility relation $R \in \{\leq, <\}$ and defined by the PDL program $\pi(R)$ using only tests, composition and union. Notice that the formula $[\sigma][R]\varphi$ evaluated at a world $w \in W$ of a model $M = \langle W, R, v\rangle$ has the same truth-value of the formula $[R]\varphi$ evaluated at w in the model $M' = \langle W, R', v\rangle$, in which $R' = \pi(R)$, i.e. R' is the result of applying the program π on the relation R. As such, we can obtain a set of reduction axioms for formulas of the type $[\sigma]\psi$ by means of the logic without the dynamic modality $[\sigma]$.

This reduction axioms can be obtained by rewriting a formula $[\sigma][R]\varphi$ as to push the dynamic modality $[\sigma]$ into the formula until it can be eliminated. Since σ is defined by a PDL program $\pi(R)$ using only tests, composition, union and the relation R, we can apply the following equivalences to compute this.

$$[\sigma][R]\varphi \leftrightarrow [\pi(R)][\sigma]\varphi$$

Applying the usual PDL axioms $[\pi \cup \sigma]\varphi \leftrightarrow [\pi]\varphi \wedge [\sigma]\varphi$, $[\pi;\sigma]\varphi \leftrightarrow [\sigma][\pi]\varphi$ and $[?\varphi]\psi \leftrightarrow \varphi \rightarrow \psi$ we can derive the desired reduction axioms. □

We wish to remind the reader not familiar with PDL the semantics of the programs. Let $M = \langle W, \leq, v\rangle$ be a preference model, we define the semantics of a PDL program π over M, denoted by $[\![\pi]\!]_M$, as the relation:

$$\begin{array}{lcl}
[\![\leq]\!]_M & = & \leq \\
[\![<]\!]_M & = & < \\
[\![\top]\!]_M & = & W^2 \\
[\![?\varphi]\!]_M & = & \{\langle w, w'\rangle \in W^2 \mid M, w \vDash \varphi \text{ and } M, w' \vDash \varphi\} \\
[\![\pi;\sigma]\!]_M & = & \{\langle w, w'\rangle \in W^2 \mid \exists w1 \in W \ s.t \ \langle w, w_1\rangle \in [\![\pi]\!]_M \text{ and } \langle w_1, w'\rangle \in [\![\sigma]\!]_M\} \\
[\![\pi \cup \sigma]\!]_M & = & [\![\pi]\!]_M \cup [\![\sigma]\!]_M
\end{array}$$

4.4.1 Public Announcement

The first operation we present is the well-known public announcement of Plaza (2007). A public announcement of φ is a truthful and knowledge increasing announcement of φ, i.e. it results in the agent coming to know that φ. Formally, a public announcement results in the agent to consider only those worlds in which the announcement is satisfied, i.e. it amounts to remove from the model each and every world not satisfying the announcement of φ.

Definition 4.23 *(GIRARD, 2008) Let $M = \langle W, \leq, v\rangle$ be a preference model and φ a formula of $\mathscr{L}_0$. We say the model $M_{!\varphi} = \langle W_{!\varphi}, \leq_{!\varphi}, v_{!\varphi}\rangle$ is the result of public announcement of φ in M, where:*

$$\begin{array}{lcl}
W_{!\varphi} & = & \{w \in W \mid M, w \vDash \varphi\} \\
\leq_{!\varphi} & = & \leq \cap (W^2_{!\varphi}) \\
v_{!\varphi}(p) & = & v(p) \cap W_{!\varphi}
\end{array}$$

We can now introduce the modality $[!\varphi]$ in the language of $\mathscr{L}_{\leq}(P)$, where $[!\varphi]\psi$ is read as "after the public announcement of φ, ψ holds".

Definition 4.24 *Let $M = \langle W, \leq, v\rangle$ be a preference model, $w \in W$, φ a formula of $\mathscr{L}_0$:*

$$M, w \vDash [!\varphi]\psi \qquad \textit{iff} \qquad M, w \vDash \varphi \textit{ implies } M_{!\varphi}, w \vDash \psi$$

Public announcements cannot be represented in the framework of the basic Propositional Dynamic Logic, since the operation of public announcement changes not only the accessibility relation, but also the domain of the structure itself. As such, we cannot derive an axiomatization using Fatc 4.22. Nevertheless, a sound and complete axiomatization for the Preference Logic extended with public announcements has been given in the literature by Girard (2008) and others.

Proposition 4.25 *(GIRARD, 2008) Preference Logic extended with public announcements is completely axiomatized by the axioms depicted in Figure 4.1, extended with the reduction axioms depicted in Figure 4.3 and the modus ponens and necessitation rules for all modalities.*

Figure 4.3 – Reduction axioms for public announcement

$$
\begin{array}{lcl}
[!\varphi]p & \leftrightarrow & \varphi \rightarrow p \\
[!\varphi]\neg\psi & \leftrightarrow & \varphi \rightarrow \neg[!\varphi]\psi \\
[!\varphi](\psi \wedge \xi) & \leftrightarrow & [!\varphi]\psi \wedge [!\varphi]\xi \\
[!\varphi]A\psi & \leftrightarrow & \varphi \rightarrow A([!\varphi]\psi) \\
[!\varphi][\leq]\psi & \leftrightarrow & \varphi \rightarrow [\leq][!\varphi]\psi \\
[!\varphi][<]\psi & \leftrightarrow & \varphi \rightarrow [<][!\varphi]\psi
\end{array}
$$

Source: the author

In order to provide a priority graph equivalent to the preference model resulting from public announcement, we need the notion of graph restriction. Graph restriction was first introduce by Van Benthem, Grossi and Liu (2014) in the context of Dynamic Deontic Logic.

Definition 4.26 *(VAN BENTHEM; GROSSI; LIU, 2014) Let $\mathcal{G} = \langle \Phi, \prec \rangle$ be a P-graph and φ a propositional formula. The restriction of $\mathcal{G}$ by φ is the graph $\mathcal{G}^{\varphi} = \langle \Phi^{\varphi}, \prec^{\varphi} \rangle$ where:*

- $\Phi^{\varphi} = \{\varphi \wedge \psi \mid \psi \in \Phi\}$
- $\prec^{\varphi} = \{\langle \varphi \wedge \psi, \varphi \wedge \psi' \rangle \mid \langle \psi, \psi' \rangle \in \prec\}$

A graph restriction, informally, consists of selecting in the induced model only those worlds satisfying a certain formula φ. With graph restriction, Van Benthem, Grossi and Liu (2014) show that public announcement can be equivalently defined with priority graphs.

Theorem 4.27 *(VAN BENTHEM; GROSSI; LIU, 2014) Let M be a preference model induced by a P-graph $\mathcal{G}$ and φ a propositional formula. If $M_{!\varphi}$ is the result of the public announcement of φ in M, then $M_{!\varphi}$ is induced by $\mathcal{G}^{\varphi}$, the restriction of $\mathcal{G}$ by φ. In other words, the diagram in Figure 4.4 commutes.*

Figure 4.4 – Harmony for Public Announcements.

$$\begin{array}{ccc} \mathscr{G} & \xrightarrow{(.)^{\varphi}} & \mathscr{G}^{\varphi} \\ {\scriptstyle \leq_{\mathscr{G}}}\downarrow & & \downarrow{\scriptstyle \leq_{\mathscr{G}^{\varphi}}} \\ M & \xrightarrow{!\varphi} & M_{!\varphi} \end{array}$$

Source: (VAN BENTHEM; GROSSI; LIU, 2014)

Theorem 4.27 is an exceptional result connecting P-graphs and preference models. In essence, it states that, considering only Public Announcements, P-graphs and preference models are equivalent encodings of preferences both on the static and dynamic perspectives, i.e. P-graphs are effective ways to represent preference models, and transformations on P-graphs are effective ways to represent semantic operations on preference models.

Using Figure 4.4 as a guide, Theorem 4.27 states that, given a p-graph $\mathscr{G}$ that induces a model M, to achieve the model $M_{!\varphi}$, we can either go from $\mathscr{G}$ to M by the induced relation in Definition 4.14, and apply the Public Announcement definition (Definition 4.23) obtaining $M_{!\varphi}$, or we can apply graph restriction in $\mathscr{G}$ and from $\mathscr{G}^{\varphi}$ by the induced relation, we can achieve $M_{!\varphi}$

For the purpose of our study, this result has an even deeper impact. Operations such as Public Announcement are kinds of mental changing operations - Public Announcement, specifically, will be interpreted as obtaining a piece of knowledge in Chapters 5 and 6. As such, Theorem 4.27 suggests that P-graphs can be used not only to reason about agents preferences, but may actually be used as data structures to compute how agents change their minds.

4.4.2 Radical Upgrade

The radical upgrade of a model by an information φ results in a model such that all worlds satisfying φ are deemed preferable than those not satisfying it. This operation corresponds to Segerberg (1998)'s irrevocable revision.

Definition 4.28 *Let $M = \langle W, \leq, v\rangle$ be a preference model and φ a formula of $\mathscr{L}_0$. We say the model $M_{\Uparrow\varphi} = \langle W, \leq_{\Uparrow\varphi}, v\rangle$ is the result of the radical upgrade of M by φ, where*

$$\leq_{\Uparrow\varphi} = (\leq \setminus \{\langle w, w'\rangle \in W^2 \mid M, w \not\Vdash \varphi \text{ and } M, w' \Vdash \varphi\}) \cup \{\langle w, w'\rangle \in W^2 \mid M, w \Vdash \varphi \text{ and } M, w' \not\Vdash \varphi\}$$

The operation above consists of making each world satisfying φ to be strictly more preferable than those not satisfying it, while maintaining the order otherwise.

We can now introduce the modality $[\Uparrow \varphi]$ in the language of $\mathscr{L}_{\leq}$, where $[\Uparrow \varphi]\psi$ is read as "after the radical upgrade by φ, ψ holds".

Definition 4.29 *Let $M = \langle W, \leq, v\rangle$ be a preference model, $w \in W$ and φ a formula of $\mathscr{L}_0$*

$$M, w \vDash [\Uparrow \varphi]\psi \qquad \textit{iff} \qquad M_{\Uparrow\varphi}, w \vDash \psi$$

The radical upgrade can be represented by a PDL program below.

$$\leq_{\Uparrow\varphi} := (?\varphi; \leq; ?\varphi) \cup (?\neg\varphi; \leq; ?\neg\varphi) \cup (?\varphi; \top; ?\neg\varphi) \tag{4.3}$$

The program above can be read as *after the radical upgrade of $\leq$ by φ, the relation $\leq_{\Uparrow\varphi}$ is composed by the pairs $\langle w, w'\rangle$ s.t. $w \vDash \varphi$, $w \leq w'$* and *$w' \vDash \varphi$*; by the pairs $\langle w, w'\rangle$ *s.t. $w \vDash \neg\varphi$*, *$w \leq w'$* and *$w' \vDash \neg\varphi$*; and by the pairs $\langle w, w'\rangle$ *s.t. $w \vDash \varphi$* and *$w' \vDash \neg\varphi$*. In the expression 4.3, the terms $(?\varphi; \leq; ?\varphi)$ and $(?\neg\varphi; \leq; ?\neg\varphi)$ correspond to preserving the relation within the sets $[\![\varphi]\!]$ and $[\![\neg\varphi]\!]$, respectively. The third term corresponds to including a link from all φ worlds to all $\neg\varphi$ worlds in the accessibility relation. Since no other links are preserved, all links from $\neg\varphi$ worlds to φ worlds are discarded.

From that, a set of reduction axioms for Preference Logic augmented with radical upgrade can be obtained by the technique described in Fact 4.22.

Proposition 4.30 *(LIU, 2011) Preference Logic extended with radical upgrade is completely axiomatized by the axioms presented in Figure 4.1 extended by the reduction axioms depicted in Figure 4.5 and the modus ponens and necessitation rules for all modalities.*

Proof: For sake of clarity, we will demonstrate how to apply Fact 4.22 to achieve the reduction axiom for the formula $[\Uparrow \varphi][\leq]\psi$.

$$\begin{aligned}
[\Uparrow \varphi][\leq]\psi &\leftrightarrow [\Uparrow (\leq)][\Uparrow \varphi]\psi \\
&\leftrightarrow [(?\varphi; \leq; ?\varphi) \cup (?\neg\varphi; \leq; ?\neg\varphi) \cup (?\varphi; \top; ?\neg\varphi)][\Uparrow \varphi]\psi \\
&\leftrightarrow [(?\varphi; \leq; ?\varphi)][\Uparrow \varphi]\psi \wedge [(?\neg\varphi; \leq; ?\neg\varphi)][\Uparrow \varphi]\psi \wedge [(?\varphi; \top; ?\neg\varphi)][\Uparrow \varphi]\psi \\
&\leftrightarrow [?\varphi][(?\varphi; \leq)][\Uparrow \varphi]\psi \wedge [?\neg\varphi][(?\neg\varphi; \leq)][\Uparrow \varphi]\psi \wedge [?\neg\varphi][(?\varphi; \top)][\Uparrow \varphi]\psi \\
&\leftrightarrow (\varphi \rightarrow [(?\varphi; \leq)][\Uparrow \varphi]\psi) \wedge (\neg\varphi \rightarrow [(?\neg\varphi; \leq)][\Uparrow \varphi]\psi) \wedge (\neg\varphi \rightarrow [(?\varphi; \top)][\Uparrow \varphi]\psi) \\
&\leftrightarrow (\varphi \rightarrow [\leq]([?\varphi][\Uparrow \varphi]\psi)) \wedge (\neg\varphi \rightarrow [\leq]([?\neg\varphi][\Uparrow \varphi]\psi)) \wedge (\neg\varphi \rightarrow A([?\varphi][\Uparrow \varphi]\psi)) \\
&\leftrightarrow (\varphi \rightarrow [\leq](\varphi \rightarrow [\Uparrow \varphi]\psi)) \wedge (\neg\varphi \rightarrow [\leq](\neg\varphi \rightarrow [\Uparrow \varphi]\psi)) \wedge (\neg\varphi \rightarrow A(\varphi \rightarrow [\Uparrow \varphi]\psi)) \\
&\leftrightarrow \varphi \rightarrow [\leq](\varphi \rightarrow [\Uparrow \varphi]\psi) \wedge \neg\varphi \rightarrow (A(\varphi \rightarrow [\Uparrow \varphi]\psi) \wedge [\leq](\neg\varphi \rightarrow [\Uparrow \varphi]\psi))
\end{aligned}$$

Figure 4.5 – Reduction axioms for the radical upgrade

$$
\begin{array}{lcl}
[\Uparrow \varphi]p & \leftrightarrow & p \\
[\Uparrow \varphi]\neg\psi & \leftrightarrow & \neg[\Uparrow \varphi]\psi \\
[\Uparrow \varphi](\psi \wedge \xi) & \leftrightarrow & [\Uparrow \varphi]\psi \wedge [\Uparrow \varphi]\xi \\
[\Uparrow \varphi]A\psi & \leftrightarrow & A([\Uparrow \varphi]\psi) \\
[\Uparrow \varphi][\leq]\psi & \leftrightarrow & \varphi \rightarrow [\leq](\varphi \rightarrow [\Uparrow \varphi]\psi) \wedge \neg\varphi \rightarrow (A(\varphi \rightarrow [\Uparrow \varphi]\psi) \wedge [\leq](\neg\varphi \rightarrow [\Uparrow \varphi]\psi)) \\
[\Uparrow \varphi][<]\psi & \leftrightarrow & \varphi \rightarrow [<](\varphi \rightarrow [\Uparrow \varphi]\psi) \wedge \neg\varphi \rightarrow (A(\varphi \rightarrow [\Uparrow \varphi]\psi) \wedge [<](\neg\varphi \rightarrow [\Uparrow \varphi]\psi))
\end{array}
$$

Source: the author

□

Inspired by the work of Andréka, Ryan and Schobbens (2002) on algebraic combinations of preference relations, Liu (2011) proposes the operation of graph prefixing.

Definition 4.31 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ and $\mathscr{G}' = \langle \Phi', \prec' \rangle$ be a P-Graphs, with $\Phi \cap \Phi' = \emptyset$. The prefixing of graph $\mathscr{G}'$ by graph $\mathscr{G}$ is the P-Graph $\mathscr{G};\mathscr{G}' = \langle \Phi \cup \Phi', \prec^{\mathscr{G};\mathscr{G}'} \rangle$ where:*

$$\prec^{\mathscr{G};\mathscr{G}'} = \prec \cup \prec' \cup \{\langle \varphi, \psi \rangle \mid \varphi \in \Phi \text{ and } \psi \in \Phi'\}$$

With this operation, Van Benthem, Grossi and Liu (2014) prove that radical upgrade can be encoded by transformations on P-graphs.

Theorem 4.32 *(VAN BENTHEM; GROSSI; LIU, 2014) Let M be a preference model induced by a P-graph $\mathscr{G}$ and φ a propositional formula. The model $M_{\Uparrow\varphi}$ is induced by the graph $\overline{\varphi};\mathscr{G}$, where $\overline{\varphi} = \langle \{\varphi\}, \emptyset \rangle$ is the singleton P-graph. In other words, the diagram in Figure 4.6 commutes.*

Figure 4.6 – Harmony for radical upgrade.

$$
\begin{array}{ccc}
\mathscr{G} & \xrightarrow{\ ;\ } & \overline{\varphi};\mathscr{G} \\
{\scriptstyle \leq_{\mathscr{G}}}\downarrow & & \downarrow{\scriptstyle \leq_{\overline{\varphi};\mathscr{G}}} \\
M & \xrightarrow{\Uparrow\varphi} & M_{\Uparrow\varphi}
\end{array}
$$

Source: (LIU, 2011)

4.4.3 Public Suggestion

Van Benthem and Liu (2007) introduce the operation of public suggestion. Taking a suggestion of φ is a less radical way of upgrading one's preferences. Semantically, taking a suggestion differs from radical revision by not introducing new links between worlds in the preference relation, only removing undesirable links of the relation. As such, this operation can also be called link-cutting upgrade.

Definition 4.33 *Let $M = \langle W, \leq, v\rangle$ be a preference model and φ a formula of $\mathscr{L}_0$. We say the model $M_{\#\varphi} = \langle W, \leq_{\#\varphi}, v\rangle$ is the result of the upgrade of M by suggestion φ, where:*

$$\leq_{\sharp\varphi} = \leq \setminus \{\langle w, w'\rangle \mid w \vDash \neg\varphi \text{ and } w' \vDash \varphi\}$$

This operation maintains all links corresponds from φ worlds to φ worlds, all links from $\neg\varphi$ worlds to $\neg\varphi$ worlds and all links from φ worlds to $\neg\varphi$ worlds, deleting all links from $\neg\varphi$ worlds to φ worlds.

We can now introduce the modality $[\#\varphi]$ in the language of $\mathscr{L}_0$, where $[\#\varphi]\psi$ is read as "after the suggestion of φ, ψ holds".

Definition 4.34 *Let $M = \langle W, \leq, v\rangle$ be a preference model, $w \in W$ and φ a formula of $\mathscr{L}_0$*

$$M, w \vDash [\#\varphi]\psi \qquad \textit{iff} \qquad M_{\#\varphi}, w \vDash \psi$$

The public suggestion operation can be represented by a PDL program bellow, introduced by Van Benthem and Liu (2007).

$$\leq_{\#\varphi} := (?\varphi; \leq; ?\varphi) \cup (?\neg\varphi; \leq; ?\neg\varphi) \cup (?\varphi; \leq; ?\neg\varphi) \tag{4.4}$$

The program above can be read as *after the upgrade of $\leq$ by public suggestion of φ, the relation $\leq_{\#\varphi}$ is composed by the pairs $\langle w, w'\rangle$ s.t. $w \vDash \varphi$, $w \leq w'$* and *$w' \vDash \varphi$*; by the pairs *$\langle w, w'\rangle$ s.t. $w \vDash \neg\varphi$, $w \leq w'$* and *$w' \vDash \neg\varphi$*; and by the pairs *$\langle w, w'\rangle$ s.t. $w \vDash \varphi$, $w \leq w'$* and *$w' \vDash \neg\varphi$*. In the expression 4.4, the terms $(?\varphi; \leq; ?\varphi)$ and $(?\neg\varphi; \leq; ?\neg\varphi)$ to preserving the relation within the sets $[\![\varphi]\!]$ and $[\![\neg\varphi]\!]$, respectively. The third term corresponds to the case in which $w \in [\![\varphi]\!]$ and $w' \notin [\![\varphi]\!]$.

From this PDL codification, applying Fact 4.22, we obtain the reduction axioms depicted in Figure 4.7. With that, we can provide an axiomatization for the logic of public suggestions, i.e. Preference Logic extended with public suggestions.

Proposition 4.35 *(VAN BENTHEM; LIU, 2007) Preference Logic extended with public suggestions is axiomatized by the axioms presented in Figure 4.1 increased by the reduction axioms depicted in Figure 4.7 and the modus ponens and necessitation rules for all modalities.*

Figure 4.7 – Reduction axioms for public suggestion

$$
\begin{array}{lcl}
[\#\varphi]p & \leftrightarrow & p \\
[\#\varphi]\neg\psi & \leftrightarrow & \neg[\#\varphi]\psi \\
[\#\varphi](\psi\wedge\xi) & \leftrightarrow & [\#\varphi]\psi\wedge[\#\varphi]\xi \\
[\#\varphi]A\psi & \leftrightarrow & A[\#\varphi]\psi \\
[\#\varphi][\leq]\psi & \leftrightarrow & \varphi\rightarrow[\leq](\varphi\rightarrow[\#\varphi]\psi)\wedge\neg\varphi\rightarrow[\leq][\#\varphi]\psi \\
[\#\varphi][<]\psi & \leftrightarrow & \varphi\rightarrow[<](\varphi\rightarrow[\#\varphi]\psi)\wedge\neg\varphi\rightarrow[<][\#\varphi]\psi
\end{array}
$$

Source: the author

The public suggestion operation can be represented as a syntactic transformation in a graph known as parallel composition, defined below.

Definition 4.36 *(VAN BENTHEM; GROSSI; LIU, 2014) Let $\mathcal{G}=\langle\Phi,\prec\rangle$ and $\mathcal{G}'=\langle\Phi',\prec'\rangle$ be P-Graphs. The parallel composition of graphs $\mathcal{G}$ and $\mathcal{G}'$ is the P-Graph $\mathcal{G}\parallel\mathcal{G}'=\langle\Phi\cup\Phi',\prec\cup\prec'\rangle$.*

From the work of Andréka, Ryan and Schobbens (2002), however, we know that the parallel composition of priority graphs is equivalent to the intersection of the induced preference relations. As such, we can construct an encoding of public suggestion by means of parallel composition without much effort.

Theorem 4.37 *(VAN BENTHEM; GROSSI; LIU, 2014) Let M be a preference model induced by a P-graph $\mathcal{G}$ and φ a propositional formula. The model $M_{\#\varphi}$ is induced by the graph $\overline{\varphi}\parallel\mathcal{G}$, where $\overline{\varphi}=\langle\{\varphi\},\emptyset\rangle$ is the singleton P-graph. In other words, the diagram in Figure 4.8 commutes.*

Figure 4.8 – Harmony for public suggestion.

$$
\begin{array}{ccc}
\mathcal{G} & \xrightarrow{\ \parallel\ } & \overline{\varphi}\parallel\mathcal{G} \\
{\scriptstyle\leq_{\mathcal{G}}}\downarrow & & \downarrow{\scriptstyle\leq_{\overline{\varphi}\parallel\mathcal{G}}} \\
M & \xrightarrow{\ \#\varphi\ } & M_{\#\varphi}
\end{array}
$$

Source: (LIU, 2011)

4.5 Dynamifying preference logic: contraction operators

We will investigate the addition of contraction operations to the Preference Logic. Different from the results presented in Section 4.4, the results in this section are original contributions of our work.

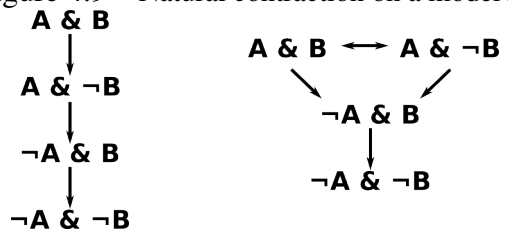
We investigate three contraction operations from the literature in Iterated Belief Contraction: Natural Contraction, Moderate Contraction and Lexicographic Contraction. These operations are three of the best known operators in the literature, as well as the ones satisfying some important properties in the context of Belief Revision Theory, such as the (generalized) Levi Identity and the Principled Factored Insertion (RAMACHANDRAN; NAYAK; ORGUN, 2012). While these properties have no direct implication on our work, they suggest the adequacy of the operations we chose to study.

As the operations we study in this section are original contributions, in the sense that we propose a codification into Dynamic Preference Logic of the operations proposed in the area of Belief Revision Theory, we change our presentation in regards to the style adopted in the last section. In this section, we chose to present the axioms or postulates that characterize these operations in the area of iterated belief revision and, based on the axioms, we then propose a definition of these operations by means of preference models. We believe this separation is important both to make explicit our contribution and to point out when and how our definition may differ with the originals.

As before, we employ the methodology sketched by Van Benthem and Liu (2007) to study dynamic operators in Dynamic Epistemic Logic, i.e. use the PDL codification of the operations and Fact 4.22 to provide reduction axioms for them. Additionally, we show that in general some of these operations are not harmonic, i.e. they are not definable by means of transformations on P-graphs. This provides a negative answer to Liu (2011)'s question of whether any operation closed over preference models that can be defined by means of PDL programs without iteration is harmonic.

This result has a methodological implication that there is no method to automatically generate a graph transformation from a PDL program without iteration that preserves preference models.

Figure 4.9 – Natural contraction on a model M



Source: the author

4.5.1 Natural Contraction

Natural contraction is a conservative contraction operation, in the sense that it aims to achieve some form of "minimal change" in the belief state, meaning that the preference relation is changed only in regards to some worlds, namely the minimal worlds not satisfying the property to be contracted. It is, in a way, a dual operation for the well-known natural revision proposed by Boutilier (1993).

Given a preference relation $\leq$ over possible worlds, its natural contraction by a sentence φ, represented by $\leq_{\downarrow\varphi}$, is defined in the work of Ramachandran, Nayak and Orgun (2012) by the following axioms:

NC1: If $w_1 \in Min_{\leq}W$ or $w_1 \in Min_{\leq}[\![\neg\varphi]\!]$, then $w_1 \leq_{\downarrow\varphi} w_2$ for any $w_2 \in W$.

NC2: If $w_1, w_2 \notin Min_{\leq}W$ and $w_1, w_2 \notin Min_{\leq}[\![\neg\varphi]\!]$ then $w_1 \leq_{\downarrow\varphi} w_2$ if and only if $w_1 \leq w_2$.

NC1 says that the most preferred worlds that don't satisfy φ are promoted to the most preferred worlds in the resulting model and NC2 states that all else remains the same.

To make it more clear, let's examine Figure 4.9. The graph to the left represents a model M, which is constructed by taking the reflexive and transitive closure of the represented relation, and the graph to the right represents the natural contraction of M by the sentence B. Notice that the natural contraction corresponds to take the minimal $\neg B$-worlds and making them minimal elements of the model.

Based on the axioms proposed to characterize Natural Contraction in the literature of Iterated Belief Revision, we define this operation by means of a transformation on preference models.

Definition 4.38 *Let $M = \langle W, \leq, v\rangle$ be a preference model and φ a formula of $\mathscr{L}_0$. We say the model $M_{\downarrow\varphi} = \langle W, \leq_{\downarrow\varphi}, v\rangle$ is the natural contraction of M by φ, where:*

$$w \leq_{\downarrow\varphi} w' \text{ iff } \begin{cases} w \in Min_{\leq} W \text{ or} \\ w \in Min_{\leq}[\![\neg\varphi]\!]_M \text{ or} \\ w \leq w' \text{ and } w' \notin Min_{\leq}[\![\neg\varphi]\!]_M \end{cases}$$

Natural contraction is our first dynamic operation that can only be defined for preference models which are well-founded. The dependence of the well-founded semantics stems from the fact that to properly define Natural Contraction we must guarantee that the set $Min_{\leq}[\![\neg\varphi]\!]$ is non-empty.

Since this operation has never before, to our knowledge, been defined on well-founded preference models, we need to investigate if the resulting model is a well-founded preference models, i.e. if the operation is closed over these models. In fact, the class of reflexive, transitive and well-founded frames is closed under the above transformation, i.e. the natural contraction preserves preference models. The reflexivity is derived by the fact that the order is preserved within the sets $[\![\varphi]\!]_M$ and $[\![\neg\varphi]\!]_M$. Transitivity follows by the fact only the minimal worlds in $[\![\neg\varphi]\!]_M$ have their positions in the preference relation altered and they become minimal. Well-foundedness follows from the fact that no infinite descending chain is created.

We can now introduce new modality $[\downarrow \varphi]$ in the logic representing the contraction of the model by φ.

Definition 4.39 *Let $M = \langle W, \leq, v\rangle$ be a preference model, $w \in W$ and φ a formula of $\mathscr{L}_0$*

$$M, w \vDash [\downarrow \varphi]\psi \qquad \textit{iff} \qquad M_{\downarrow\varphi}, w \vDash \psi$$

To provide a PDL representation of Natural Contraction, we need to provide a sub-program for each clause in Definition 4.38. The first term requires that if $w \in Min_{\leq} W$ than $w \leq_{\downarrow\varphi} w'$ for any w'. Well, $w \in Min_{\leq} W$ may be represented by the program $?\mu\top$ and the condition for any w' can be represented by the program $\top$, taking the composition of both programs $(?\mu\top;\top)$ we achieve the desired condition. The second clause is similar only changing the program $?\mu\top$ to $?\mu\neg\varphi$. Finally, in the third clause $w \leq w'$ can be represented by the program $\leq$ while $w' \notin Min_{\leq}[\![\neg\varphi]\!]$ can be represented by $?\neg\mu\neg\varphi$. Composing both programs we achieve $(\leq;?\neg\mu\neg\varphi)$. The natural contraction can, thus, be represented by the following PDL program:

$$\downarrow_{\varphi}(\leq) := (?\mu\top;\top) \cup (?\mu\neg\varphi;\top) \cup (\leq;?\neg\mu\neg\varphi) \tag{4.5}$$

From this PDL representation we can easily provide an axiomatization for the logic extended with the $[\downarrow \varphi]$ modality, applying Fact 4.22.

Proposition 4.40 *The Preference Logic extended with Natural Contraction is soundly and completely axiomatized by the reduction axioms presented in Figure 4.10 added to the axioms provided in Figure 4.1 for the Preference Logic.*

Proof: Immediate from Fact 4.22 and expression 4.5. □

Figure 4.10 – Reduction axioms for Natural Contraction

$$
\begin{array}{lcl}
[\downarrow \varphi]p & \leftrightarrow & p \\
[\downarrow \varphi]\neg\psi & \leftrightarrow & \neg[\downarrow \varphi]\psi \\
[\downarrow \varphi](\xi \wedge \psi) & \leftrightarrow & [\downarrow \varphi]\xi \wedge [\downarrow \varphi]\psi \\
[\downarrow \varphi]A\psi & \leftrightarrow & A[\downarrow \varphi]\psi \\
[\downarrow \varphi][\leq]\psi & \leftrightarrow & A(\mu\top \rightarrow [\downarrow \varphi]\psi) \wedge \\
 & & A(\mu\neg\varphi \rightarrow [\downarrow \varphi]\psi) \wedge \\
 & & \neg\mu\neg\varphi \rightarrow [\leq]([\downarrow \varphi]\psi) \\
[\downarrow \varphi][<]\psi & \leftrightarrow & \neg\mu\neg\varphi \rightarrow [<]([\downarrow \varphi]\psi)
\end{array}
$$

Source: the author

We now show that Natural contraction cannot be represented by means of priority graphs. This is a somewhat surprising result for two reasons. First, Rott (2009) has already proposed a codification of Natural Contraction by means of operations on stratified belief bases - a restricted form of priority graph. Second, because it implies that not every operation defined over preference models can be represented by transformations on priority graphs.

To see the problem with Rott (2009) codification of Natural Contraction, take the graph $\mathscr{G} = \langle \Phi = \{p,q,r\}, \prec \rangle$ where $p \prec q \prec r$ and $M = \langle 2^{\Phi}, \leq, v \rangle$ a model induced by $\mathscr{G}$, with $w \in v(u)$ iff $u \in w$ for $u \in \{p,q,r\}$. Rott (2009)'s proposal is that the Natural Contraction of q in this model can be represented by an operation that results in the graph $\mathscr{G}' = \langle \{p \vee \neg q, q \vee \neg q, r \vee \neg q, p\}, \prec' \rangle$ where $\prec' = p \vee \neg q \prec' q \vee \neg q \prec' r \vee \neg q \prec' p$.

Well, take the model $M' = \langle 2^{\{p,q,r\}}, \leq', v \rangle$ induced by $\mathscr{G}'$, with $w \in v(u)$ iff $u \in w$ for $u \in \{p,q,r\}$. In M', the worlds $p \wedge \neg q \wedge r$ and $p \wedge \neg q \wedge \neg r$ are in the same preference cluster. Natural contraction, however, would require by order preservation in $[\![\neg q]\!]$ that $p \wedge \neg q \wedge r \leq' p \wedge \neg q \wedge \neg r$. Hence $M' \neq M_{\downarrow\varphi}$.

In fact, Natural Contraction is inherently model-dependent and cannot be characterised by means of priority graphs. In the Theorem 4.41 below, we show that there are two models M_1 and M_2 that are induced by the the same P-graph $\mathscr{G}$, but the models $M_{1\downarrow\varphi}$ and $M_{2\downarrow\varphi}$ are

necessarily induced by two different P-graphs $\mathscr{G}_1$ and $\mathscr{G}_2$. As such, there is no transformation on $\mathscr{G}$ that can simultaneously describe the Natural Contraction in both M_1 and M_2, i.e. that describes the operation of Natural Contraction.

To prove that, we will require two simple conditions. The first is that the contraction is non-trivial, i.e. given a model M to be contracted by φ, all its minimal elements satisfy φ and $\neg\varphi$ is satisfiable in M. This is because, if that is not the case, the natural contraction of such model has no effects, i.e. $M_{\downarrow\varphi} = M$. Further yet, we will require that there is chain of at least two worlds in M not satisfying φ, i.e. at least two elements w, w' in M not satisfying φ s.t. $w < w'$. This is because otherwise all the elements of $[\![\neg\varphi]\!]$ would be minimal in this set.

Theorem 4.41 below expresses the fact that it is possible that contracting a formula from two models induced by the same priority graph can result in different models, in the sense that there is no priority graph that induces both transformed models. In other words, there is no syntactic operation over priority graphs such that we can construct a commutative diagram as in Figure 4.4 for Public Announcements, for instance.

Theorem 4.41 *Let $\mathscr{G}$ be a P-Graph,φ a propositional formula and $M_1 = \langle W_1, \leq_1, v_1\rangle$ a preference model induced by $\mathscr{G}$ having a chain of at least two worlds satisfying $\neg\varphi$ and such that $Min_{\leq_1} W_1 \subseteq [\![\varphi]\!]$. There is a preference model $M_2 = \langle W_2, \leq_2, v_2\rangle$ induced by $\mathscr{G}$ s.t. there is no P-graph $\mathscr{G}'$ that both $M_{1\downarrow\varphi}$ and $M_{2\downarrow\varphi}$ are preference models induced by $\mathscr{G}'$.*

Proof: Let M_1 be a preference model induced by $\mathscr{G}$. Since $M_1 = \langle W_1, \leq_1, v_1\rangle$ is a preference model and $[\![\neg\varphi]\!]_{M_1} \neq \emptyset$, then there are minimal elements in $[\![\neg\varphi]\!]_{M_1}$. Lets call one such element $w_1 \in Min_{\leq_1}[\![\neg\varphi]\!]_{M_1}$. We can construct $M_2 = \langle W_2 = W_1 \setminus Min_{\leq_1}[\![\neg\varphi]\!]_{M_1}, \leq_2, v_2\rangle$ s.t. $\leq_2$ and v_2 are the restriction of $\leq_1$ and v_1 to W_2, respectively.

Now, suppose there is a P-graph $\mathscr{G}' = \langle \Phi', \prec'\rangle$, s.t. $M_{1\downarrow\varphi} = \langle W_1, \leq_{1\downarrow\varphi}, v\rangle$ is a preference model induced by $\mathscr{G}'$. By the definition of natural contraction, we have that $M_{1\downarrow\varphi}$ is exactly like M_1, except that $Min_{\leq_{1\downarrow\varphi}} W_1 = Min_{\leq_1} W_1 \cup Min_{\leq_1}[\![\neg\varphi]\!]_{M_1}$. Take $w_1 \in Min_{\leq_1}[\![\neg\varphi]\!]$, by definition of an induced preference model, it means that for every formula $\xi \in \Phi'$ and any world $w \in Min_{\leq_1} W_1$, $M_1, w \vDash \xi$ iff $M_1, w_1 \vDash \xi$.

Let's then look at the case of $w_2 \in Min_{\leq_2}[\![\neg\varphi]\!]_{M_2}$. We know that w_2 exists since M_2 was created by removing the minimal elements of $[\![\neg\varphi]\!]_{M_1}$ from the set W_1 and, by hypothesis, $[\![\neg\varphi]\!]_{M_1}$ has a chain of at least two elements, thus not all elements of $[\![\neg\varphi]\!]_{M_1}$ are minimal. Now, if $M_{2\downarrow\varphi}$ is induced by $\mathscr{G}'$, we have that for every formula $\xi \in \Phi'$ and any world $w \in Min_{\leq_1} W_2$, $M_1, w \vDash \xi$ iff $M_1, w_2 \vDash \xi$. Since the contraction is non-trivial, we have that $Min_{\leq_1} W_2 \cap Min_{\leq_1} W_1$ is non empty. From that, we have that $M_1, w_1 \vDash \xi$ iff $M_1, w_2 \vDash \xi$ and from this we can affirm that w_2 is a minimal element in $M_{1\downarrow\varphi}$, and thus also in $[\![\neg\varphi]\!]_{M_1}$, which is a contradiction to the hypothesis that $w_1 <_1 w_2$. So $M_{2\downarrow\varphi}$ cannot be induced by $\mathscr{G}'$. □

The theorem above is a negative answer to the question posed by Liu (2011) about whether any PDL-definable transformation preserving preference models can be characterised by means of syntactic transformations in priority graphs. The root of the problem here lies in the fact that priority graphs are defined over propositional formulas only, and the notion of 'minimal element' is necessarily modal. Since Natural Contraction is defined by means of the set of minimal elements satisfying a given formula, there is no general construction that will be applicable for any induced model. Notice however that allowing modal formulas in priority graphs incurs in invalidating Theorem 4.17 by allowing graphs that have no induced model (LIU, 2011).

4.5.2 Moderate Contraction

As it has been done for natural contraction, we will extend the language of the preference logic to include a new modality $[\Downarrow \varphi]$ representing the operation of contracting a formula φ of the model by means of the moderate contraction. Moderate contraction is a less conservative form of iterated contraction, in which not only the minimal $\neg\varphi$ worlds have their preference status changed, but also of any other $\neg\varphi$ world

The moderate contraction of a preference relation $\leq$ over possible worlds by a sentence φ, represented by $\leq_{\Downarrow\varphi}$, is defined by Ramachandran, Nayak and Orgun (2012) by the following axioms:

MC1: If $w_1 \vDash \varphi$ and $w_2 \vDash \varphi$ then $w_1 \leq_{\Downarrow\varphi} w_2$ if and only if $w_1 \leq w_2$.

MC2: If $w_1 \vDash \neg\varphi$ and $w_2 \vDash \neg\varphi$ then $w_1 \leq_{\Downarrow\varphi} w_2$ if and only if $w_1 \leq w_2$.

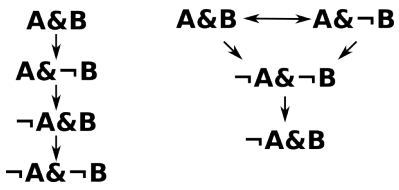
MC3: If $w_1 \vDash \varphi$, $w_1 \notin Min_{\leq}W$ and $w_2 \vDash \neg\varphi$ then $w_2 <_{\Downarrow\varphi} w_1$.

MC4: If $w_1 \in Min_{\leq}W$ or $w_1 \in Min_{\leq}[\![\neg\varphi]\!]$ then $w_1 \leq_{\Downarrow\varphi} w_2$, for any w_2.

MC1 and MC2 require that within the sets $[\![\varphi]\!]$ and $[\![\neg\varphi]\!]$ the order is preserved; MC3 requires that the worlds not satisfying φ are more preferable than any world satisfying φ that is not minimal in the original preference relation. Finally MC4 says that after the contraction the minimal elements of the resulting preference relation are exactly the original minimal elements plus the minimal elements in $[\![\neg\varphi]\!]$.

To make it more clear, let's examine Figure 4.11. The graph to the left represents a model M, which is constructed by taking the reflexive and transitive closure of the represented relation, and the graph to the right represents the natural contraction of M by the sentence B.

Figure 4.11 – Moderate contraction on a model M



Source: the author

As before, we define the moderate contraction as a transformation on models.

Definition 4.42 *Let $M = \langle W, \leq, v\rangle$ be a preference model and φ a formula of $\mathscr{L}_0$. We say the model $M_{\downdownarrows\varphi} = \langle W, \leq_{\downdownarrows\varphi}, v\rangle$ is the moderate contraction of M by φ, where:*

$$w \leq_{\downdownarrows\varphi} w' \text{ iff } \begin{cases} w \leq w' \text{ and } w, w' \in [\![\varphi]\!]_M \text{ or } w, w' \in [\![\neg\varphi]\!]_M \text{ or} \\ w \in [\![\neg\varphi]\!]_M, w' \in [\![\varphi]\!]_M \text{ and } w' \notin Min_{\leq} W \text{ or} \\ w \in Min_{\leq} W \text{ or} \\ w \in Min_{\leq} [\![\neg\varphi]\!]_M \end{cases}$$

As before, it is easy to see that the above transformation preserves preference models. We can now introduce in the language of preference logic, the modality $[\downdownarrows\ \varphi]$, as it has been done for natural contraction.

Definition 4.43 *Let $M = \langle W, \leq, v\rangle$ be a preference model, $w \in W$ and φ a formula of $\mathscr{L}_0$*

$$M, w \vDash [\downdownarrows\ \varphi]\psi \qquad \textit{iff} \qquad M_{\downdownarrows\varphi}, w \vDash \psi$$

As before, we encode this transformation in a PDL program in order to provide an axiomatization for the augmented logic.

As for Natural Contraction, the construction of the PDL representation for Moderate Contraction consists of translating each clause of Definition 4.42 as a PDL program. This operation can, thus, be represented by the following PDL program:

$$\downdownarrows\ \varphi(\leq) := \leq_{\varphi} \cup (?\neg\varphi; \top; ?\varphi \wedge \neg\mu\top) \cup (?\mu\top; \top) \cup (?\mu\neg\varphi; \top) \tag{4.6}$$

where

$$\leq_{\varphi} = (?\varphi; \leq; ?\varphi) \cup (?\neg\varphi; \leq; ?\neg\varphi)$$

With the PDL representation presented in the expression 4.6 we can provide the axiomatization below.

Proposition 4.44 *The Preference Logic extended with Moderate Contraction is soundly and completely axiomatized by the axioms provided in Figure 4.1 for the Preference Logic extended with the reduction axioms depicted in Figure 4.12 and the modus ponens and necessitation rules for all modalities.*

Proof: Immediate from Fact 4.22 and expression 4.6. □

Figure 4.12 – Reduction axioms for Moderate Contraction

$$
\begin{array}{lcl}
[\downdownarrows \varphi]p & \leftrightarrow & p \\
[\downdownarrows \varphi]\neg\psi & \leftrightarrow & \neg[\downdownarrows \varphi]\psi \\
[\downdownarrows \varphi](\xi \wedge \psi) & \leftrightarrow & [\downdownarrows \varphi]\xi \wedge [\downdownarrows \varphi]\psi \\
[\downdownarrows \varphi]A\psi & \leftrightarrow & A[\downdownarrows \varphi]\psi \\
[\downdownarrows \varphi][\leq]\psi & \leftrightarrow & (\varphi \rightarrow [\leq](\varphi \rightarrow [\downdownarrows \varphi]\psi)) \wedge \\
 & & (\neg\varphi \rightarrow [\leq](\neg\varphi \rightarrow [\downdownarrows \varphi]\psi)) \wedge \\
 & & (\varphi \wedge \neg\mu\top \rightarrow A(\neg\varphi \rightarrow [\downdownarrows \varphi]\psi)) \wedge \\
 & & A(\mu\top \rightarrow [\downdownarrows \varphi]\psi) \wedge \\
 & & A(\neg\varphi \rightarrow [\downdownarrows \varphi]\psi)) \\
[\downdownarrows \varphi][<]\psi & \leftrightarrow & (\varphi \rightarrow [<](\varphi \rightarrow [\downdownarrows \varphi]\psi)) \wedge \\
 & & (\neg\varphi \rightarrow [<](\neg\varphi \rightarrow [\downdownarrows \varphi]\psi)) \wedge \\
 & & (\varphi \wedge \neg\mu\top \rightarrow A(\neg\varphi \rightarrow [\downdownarrows \varphi]\psi)) \wedge
\end{array}
$$

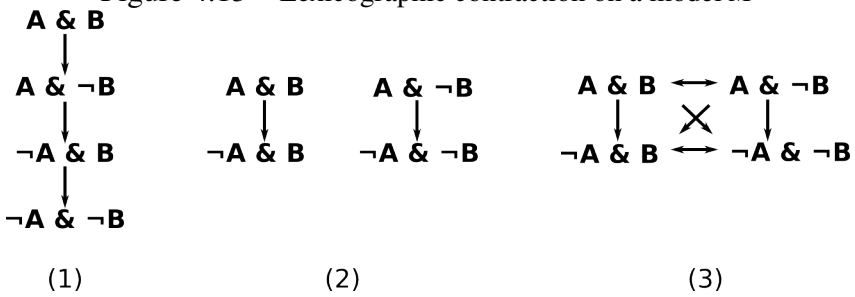
Source: the author

As for Natural Contraction we can also prove that Moderate Contraction is inherently model-dependent and cannot be characterised by means of priority graphs.

Theorem 4.45 *Let $\mathscr{G}$ be a P-Graph,φ a propositional formula and $M_1 = \langle W_1, \leq_1, v_1\rangle$ a preference model induced by $\mathscr{G}$ having a chain of at least two worlds satisfying $\neg\varphi$ and such that $Min_{\leq_1}W_1 \subseteq [\![\varphi]\!]$. There is a preference model $M_2 = \langle W_2, \leq_2, v_2\rangle$ induced by $\mathscr{G}$ s.t. there is no P-graph $\mathscr{G}'$ that both $M_{1\downdownarrows\varphi}$ and $M_{2\downdownarrows\varphi}$ are preference models induced by $\mathscr{G}'$.*

The proof of Theorem 4.45 is similar to that of Theorem 4.41. In fact, we point out that for any model M, $M_{\downdownarrows\varphi} = (M_{\Uparrow\neg\varphi})_{\downarrow\neg\varphi}$.

Figure 4.13 – Lexicographic contraction on a model M



Source: the author

4.5.3 Lexicographic Contraction

The lexicographic contraction was introduced by Nayak et al. (2006), as a product of the generalization of the well-known Harper identity (ALCHOURRÓN; GÄRDENFORS; MAKINSON, 1985). Despite the desirable properties it satisfies, in the context of Belief Revision Theory, this operation presents difficulties in characterization as will be evident in Proposition 4.50 below. This is because it is defined by means of complete chains of worlds in a model - which encodes a great deal of information about the preference relation.

The lexicographic contraction of a preference relation $\leq$ over possible worlds by a sentence φ, represented by $\leq_{\Downarrow\varphi}$, is defined by Ramachandran, Nayak and Orgun (2012) by the following axioms:

LC1: If $w_1 \vDash \varphi$ and $w_2 \vDash \varphi$ then $w_1 \leq_{\Downarrow\varphi} w_2$ if and only if $w_1 \leq w_2$.

LC2: If $w_1 \vDash \neg\varphi$ and $w_2 \vDash \neg\varphi$ then $w_1 \leq_{\Downarrow\varphi} w_2$ if and only if $w_1 \leq w_2$.

LC3: Let ξ be a member of $\{\varphi, \neg\varphi\}$ and $\overline{\xi}$ the other. If $w_1 \vDash \xi$ and $w_2 \vDash \overline{\xi}$, then $w_1 \leq_{\Downarrow\varphi} w_2$ iff the length of a complete chain of worlds in $[\![\xi]\!]$ which ends in w_1 is less than or equal to the length of a complete chain of worlds in $[\![\overline{\xi}]\!]$ which ends in w_2.

As for the MC1 and MC2, LC1 and LC2 require order preservation in $[\![\varphi]\!]$ and $[\![\neg\varphi]\!]$. LC3 requires that the preference relation will be computed by lexicographically joining the equivalence classes in $[\![\varphi]\!]$ and $[\![\neg\varphi]\!]$, regarding the preference relation $\leq$.

To make it more clear, let's examine the example depicted in Figure 4.13. As before, if there is an edge from world w to world w', then $w \leq w'$. In Figure 4.13, (1) represents a model M, which can be constructed by taking the reflexive and transitive closure of the represented relation, (2) the chains of worlds in $[\![B]\!]$ (left) and $[\![\neg B]\!]$ (right), and (3) the lexicographic contraction of M by the sentence B, which corresponds to joining the nodes with same depth in the chains depicted in (2) into a same $\leq$-cluster.

To specify the lexicographic contraction we need a way to represent that there is a chain of worlds in $[\![\varphi]\!]$ of length i. We will define a formula $dg_\varphi(i)$ below to represent this notion.

Definition 4.46 *Let $M = \langle W, \leq, v\rangle$ be a preference model, φ a formula of $\mathscr{L}_0$ and $i \in \mathbb{N}$ a natural number. We define the formula $dg_\varphi(i)$ as*

$$dg_\varphi(i) = \begin{cases} \varphi & \text{if } i = 1 \\ \varphi \wedge \langle < \rangle dg_\varphi(i-1) & \text{if } i > 1 \end{cases}$$

It is easy to see that $dg_\varphi(i)$ encodes the notion of existence of a chain of worlds in $[\![\varphi]\!]$ of length i[8], since for each i, $M, w \vDash dg_\varphi(i)$ means that $w \in [\![\varphi]\!]$ and there is a world w', s.t. $w' < w$ and $M, w' \vDash dg_\varphi(i-1)$.

Lemma 4.47 *Let $M = \langle W, \leq, v\rangle$ be a preference model, φ a formula of $\mathscr{L}_0$ and $w \in W$. $M, w \vDash dg_\varphi(i), i > 1$ iff there is a chain of worlds of $w_1 < w_2 \ldots < w_i$, such that, $w_j \in [\![\varphi]\!]$, for all $j = 1..i$, and $w_i = w$.*

Proof: The proof is carried by simple induction on the parameter i, noticing that $<$ is a transitive relation □

With that, we can define the lexicographic contraction as a model transformation.

Definition 4.48 *Let $M = \langle W, \leq, v\rangle$ be a preference model and φ a formula of $\mathscr{L}_0$. We say the model $M_{\Downarrow\varphi} = \langle W, \leq_{\Downarrow\varphi}, v\rangle$ is the lexicographic contraction of M by φ, where:*

$$w \leq_{\Downarrow\varphi} w' \text{ iff } \begin{cases} w \leq w' & w, w' \in [\![\varphi]\!] \\ w \leq w' & w, w' \in [\![\neg\varphi]\!] \\ w \in [\![\mu dg_\varphi(i)]\!] \text{ and } w' \in [\![\mu dg_{\neg\varphi}(j)]\!] & i \leq j \leq |W| \\ w \in [\![\mu dg_{\neg\varphi}(i)]\!] \text{ and } w \in [\![\mu dg_\varphi(j)]\!] & i \leq j \leq |W| \end{cases}$$

It is important to notice that in the Definition 4.48, we encode the condition $w \leq_{\Downarrow\varphi} w'$ iff $w \leq w'$ of Ramachandran, Nayak and Orgun (2012)'s LC3 axiom as $i < j$ in both conditions $w \in [\![\mu dg_\varphi(i)]\!]$ and $w' \in [\![\mu dg_{\neg\varphi}(j)]\!]$ and $w \in [\![\mu dg_{\neg\varphi(i)}]\!]$ and $w' \in [\![\mu dg_\varphi(j)]\!]$ for the simple fact that those authors supposed a connected preference relation, i.e. for any two worlds w, w' either $w \leq w'$ or $w' \leq w$. Since we do not require connectivity, we have to alter their definition to guarantee transitivity of the resulting relation. For the special case of connected preference relations, however, our definition is equivalent to theirs.

[8]Particularly, if $\varphi = \top$ this formula encodes the notion of the degree of a world - similar to that of Spohn (1988). Notice that only the maximal i s.t. a world satisfies $M, w \vDash dg_\top(i)$ can be thought as the degree of world w as in the framework of Spohn's ordinal conditional functions, since by transitivity $M, w \vDash dg_\top(j)$ for all $1 \leq j \leq i$.

By similar arguments as for natural and moderate contraction, it is not difficult to see that the lexicographic contraction operation preserves preference models. Then, we can now include the operation as a modality in the language of Preference Logic.

Definition 4.49 *Let $M = \langle W, \leq, v\rangle$ be a preference model, $w \in W$, and φ a formula of $\mathscr{L}_0$*

$$M, w \vDash [\Downarrow \varphi]\psi \qquad \textit{iff} \qquad M_{\Downarrow\varphi}, w \vDash \psi$$

Notice that Definition 4.48 depends on the sizes of the chains in $[\![\varphi]\!]$ and $[\![\neg\varphi]\!]$, encoded by the formulas $d_\varphi(i)$ and $d_{\neg\varphi}(i)$, respectively. To provide a finite PDL encoding of this operation, by means of the formulas $d_\varphi(i)$ and $d_{\neg\varphi}(i)$, however, we must guarantee that there is an upper bound on the number i that we must investigate.

Let $n < \infty$ be the size of the biggest $<$-chain of worlds in a model $M = \langle W, \leq, v\rangle$. To compute $M_{\Downarrow\varphi}$, according to Definition 4.48, we must then only consider the combinations of $\mu d_\varphi(i)$ and $\mu d_{\neg\varphi}(j)$, for $i, j \leq n$. If we know n beforehand, then we can provide an encoding for Lexicographic Contraction by the following PDL program:

$$\begin{aligned} \leq_{\Downarrow\varphi} := \quad & (?\varphi;\leq;?\varphi) \cup (?\neg\varphi;\leq;?\neg\varphi) \cup \\ \bigcup_{i=0}^{n}\bigcup_{j=i}^{n} \quad & \Big(((?\mu dg_\varphi(i);\top;?\mu dg_{\neg\varphi}(j)) \cup \\ & ((?\mu dg_{\neg\varphi}(i);\top;?\mu dg_\varphi(j))\Big) \end{aligned} \tag{4.7}$$

The problem now is to determine if there is an upper bound n that for any model there is no $<$-chain with size bigger then n. This upper bound, however, can't exist. To see that, notice that for any model $M = \langle W, \leq, v\rangle$ with a biggest chain of size k, we can construct a world $M' = \langle W', \leq', v'\rangle$ with with a biggest chain of size k' and $k' > k$[9].

If we can limit the models to be considered in the semantics to those having no $<$-chain of size bigger than n, however, the program 4.7 gives us a correct encoding of lexicographic contraction. Well, this property can, in fact, be expressed in our logic by the formula $\neg d_\top(n+1)$. In this case, using the PDL reduction provided by Fact 4.22, we can give a parametrized axiomatization of Preference Logic extended with lexicographic contraction, in which we consider only models with no chains of size greater then n, for any $1 \leq n < \infty$.

Proposition 4.50 *The Preference Logic extended with Lexicographic Contraction, restricted to models with no $<$-chains bigger than n, is soundly and completely axiomatized by the axioms provided in Figure 4.1 for the Preference Logic extended with the reduction axioms depicted in Figure 4.14 and the modus ponens and necessitation rules for all modalities.*

[9] This can be done, for example, taking $W' = W \times \{0,1\}$ and $\langle w,i\rangle \leq' \langle w',i'\rangle$ iff either $w \leq w'$ and $i \leq i'$ or $i < i'$.

Proof: Immediate from Fact 4.22 and Lemma 4.47 and the PDL formula 4.7. □

Figure 4.14 – Reduction axioms for Lexicographic Contraction

$$
\begin{array}{lcl}
\neg d_{\top}(n+1) & & \\
[\Downarrow \varphi]p & \leftrightarrow & p \\
[\Downarrow \varphi]\neg\psi & \leftrightarrow & \neg[\Downarrow \varphi]\psi \\
[\Downarrow \varphi](\xi \wedge \psi) & \leftrightarrow & [\Downarrow \varphi]\xi \wedge [\Downarrow \varphi]\psi \\
[\Downarrow \varphi]A\psi & \leftrightarrow & A[\Downarrow \varphi]\psi \\
[\Downarrow \varphi][\leq]\psi & \leftrightarrow & \varphi \rightarrow [\leq](\varphi \rightarrow [\Downarrow \varphi]\psi) \wedge \\
 & & \neg\varphi \rightarrow [\leq](\neg\varphi \rightarrow [\Downarrow \varphi]\psi) \wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i}^{n} \mu dg_{\varphi}(j) \rightarrow A(\mu dg_{\neg\varphi}(i) \rightarrow [\Downarrow \varphi]\psi) \wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i}^{n} \mu dg_{\neg\varphi(j)} \rightarrow A(\mu dg_{\varphi}(i) \rightarrow [\Downarrow \varphi]\psi) \\
[\Downarrow \varphi][<]\psi & \leftrightarrow & \varphi \rightarrow [<](\varphi \rightarrow [\Downarrow \varphi]\psi) \wedge \\
 & & \neg\varphi \rightarrow [<](\neg\varphi \rightarrow [\Downarrow \varphi]\psi) \wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i+1}^{n} \mu dg_{\varphi}(j) \rightarrow A(\mu dg_{\neg\varphi}(i) \rightarrow [\Downarrow \varphi]\psi) \wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i+1}^{n} \mu dg_{\neg\varphi}(j) \rightarrow A(\mu dg_{\varphi}(i) \rightarrow [\Downarrow \varphi]\psi)
\end{array}
$$

Source: the author

Notice that the axiomatization of Figure 4.14 is parametrized by the size of the biggest chain. This means we do not have a correct and complete axiomatization for the extended logic considering all possible models. While, from a logical point of view our solution is not optimal, for our pragmatic concerns of using Dynamic Preference Logic to study Agent Programming this parametrization will be enough. The reason for this is that, as we discuss in Section 4.6, in our implementation of the language we will focus on a special class of models briefly discussed in Section 4.3, namely broad models (c.f. Definition 4.20), and for these models we know the upperbound $2^P = |W|$ for the size of the biggest chain.

A codification for lexicographic contraction by means of syntactic structures has not yet been proposed and, in fact, it is not an easy one to provide. To provide such encodings, we will require some further conditions on priority graphs.

Lemma 4.51 *(LIU, 2011) Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-graph and $M = \langle W, \leq, v \rangle$ a preference model induced by $\mathscr{G}$, there is a graph $\mathscr{G}' = \langle \Phi', \prec' \rangle$ s.t. M is also induced by $\mathscr{G}'$ and for any $\xi_1, \xi_2 \in \Phi'$, if there is a $w \in W$ with $M, w \vDash \xi_1$ and $M, w \vDash \xi_2$, then $\xi_1 = \xi_2$. In other words, each priority graph has an equivalent graph whose propositions form a partition of the logical space.*

The above result means that any priority graph can be rewritten to an equivalent one in such a way that any world in a model satisfies exactly one formula in the resulting P-graph. If a P-graph $\mathscr{G}$ satisfies this condition, i.e. if its propositions form a partition of the logical space, we will say $\mathscr{G}$ is in its normal form.

To describe the transformation in the priority graph, we will need some auxiliary constructions.

The first one we present is the support of φ in a graph $\mathscr{G}$.

Definition 4.52 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph and φ a propositional formula. We define the support of φ in $\mathscr{G}$ as the graph $\mathscr{G}_\varphi = \langle \Phi_\varphi, \prec_\varphi \rangle$ where: $\Phi_\varphi = \{\xi \wedge \varphi \in \Phi \mid \xi \nvDash \neg\varphi\}$ and $\xi \wedge \varphi \prec_\varphi \xi' \wedge \varphi$ iff $\xi \prec \xi'$.*

We will usually refer to the $\mathscr{G}_\varphi$ and $\mathscr{G}_{\neg\varphi}$ as the support graphs of φ, or simply the support graphs, when the formula φ is clear. The second definition we need is that of model adequateness.

Definition 4.53 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph and $M = \langle W, \leq, v \rangle$ a preference model induced by $\mathscr{G}$. We say $\mathscr{G}$ is adequate in respect to model M iff for any $\xi \in \Phi$, there is at least a world $w \in W$, s.t. $M, w \vDash \xi$ and there is at least a world $w' \in W$, s.t. $M, w' \nvDash \xi$.*

Notice that for any P-graph and any model, we can transform the P-graph to construct a model-adequate one by removing those elements in the graph which are either not satisfied or valid in the model.

With these two definitions, we can show a very interesting result connecting support of φ in a graph and the degree of a world regarding φ. Notice that the support of φ in a graph $\mathscr{G}$ contains all information about the chains of worlds satisfying φ for models induced by $\mathscr{G}$. In other words, the support of φ in $\mathscr{G}$ is a nearly-complete description of the order in $[\![\varphi]\!]$, for any model M induced by $\mathscr{G}$. This is what we show in the following result - which can be easily proved by an induction on n.

Lemma 4.54 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph in its normal form, $M = \langle W, \leq, v \rangle$ a model induced by $\mathscr{G}$ and φ a propositional formula. If $\mathscr{G}_\varphi$ is adequate in respect to the restricted model $M_\varphi = \langle [\![\varphi]\!], \leq \cap ([\![\varphi]\!]^2), v \cap (P \times 2^{[\![\varphi]\!]}) \rangle$, for any $w \in W$, it holds that $M, w \vDash dg_\varphi(n)$ iff there is a sequence $\xi_1, \ldots, \xi_n \in G_\varphi$ s.t. $\xi_i < \xi_{i+1}$ and $M, w \vDash \xi_n$.*

The lemma above means that if the support of φ in $\mathscr{G}$ does not have irrelevant elements, than the notion of chains in the support graph corresponds to the notion of chains the the induced

model. For that to hold, all we have to guarantee is that the support graph is adequate in respect to the model.

Lemma 4.55 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph and $M = \langle W, \leq, \rangle$ a preference model induced by $\mathscr{G}$ and φ a propositional formula. We can construct a P-graph $\mathscr{G}'$, such that $\mathscr{G}'$ is adequate in respect to M and $\mathscr{G}_\varphi$ and $\mathscr{G}_{\neg\varphi}$ are adequate in respect to M_φ and $M_{\neg\varphi}$, respectively.*

Proof: Without loss of generality, we will assume $\mathscr{G}$ is in its normal form. First remove all elements of $\mathscr{G}$ that are either valid or unsatisfiable in M, obtaining the graph $\mathscr{G}_1 = \langle \Phi_1, \prec \rangle$. Notice that, since the set of propositional symbols P is finite, for any world $w \in W$, there is a propositional formula ψ_w s.t. for any propositional formula $\xi \in \mathscr{L}_0$, $M, w \vDash \xi$ iff $\psi_w \Rightarrow \xi$. In particular, take

$$\psi_w = \bigwedge \{p \in P \mid w \in v(p)\} \cup \{\neg p \mid p \in P \text{ and } w \notin v(p)\}.$$

We can now construct $\mathscr{G}' = \langle \Phi', \prec' \rangle$ s.t.

$$\Phi' = \{\xi \wedge \bigvee \{\psi_w \mid M, w \vDash \xi\}\}$$

and $\xi_1 \wedge \bigvee\{\psi_w\} \prec \xi_2 \wedge \bigvee\{\psi_w\}$ iff $\xi_1 \prec \xi_2$. By construction, $\mathscr{G}'$ is adequate in regards to M.

The key step to prove that the support graphs are adequate is to notice that for any $\xi \wedge \bigvee\{\psi_w\}\} \in \Phi'$, if $\xi \wedge \bigvee\{\psi_w\}\} \not\Rightarrow \neg\varphi$, then there is a world $w \in W$, s.t. $\psi_w \Rightarrow \varphi$, i.e. $M, w \vDash \varphi$. As such, for any $\xi \wedge \bigvee\{\psi_w\}\} \wedge \varphi$ in the support graph $\mathscr{G}'_\varphi$, there is at least a world w in M_φ s.t. $M_\varphi, w \vDash \xi \wedge \bigvee\{\psi_w\}\} \wedge \varphi$. The case for $\mathscr{G}'_{\neg\varphi}$ is similar. □

From the results above, we can see that the notion of the size of a chain in a P-graph is intimately related with the degrees of worlds in all models in which respect this P-graph is adequate. To simplify the notation, we will define the notion of the depth of a formula in the graph.

Definition 4.56 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph and $\xi \in \Phi$ a propositional formula. We define the depth of ξ in $\mathscr{G}$, denoted by $d_{\mathscr{G}}(\xi) = k$, as the size of the longest chain $\xi_1, \ldots, \xi_k$ in $\mathscr{G}$ s.t. ξ_1 is minimal in $\mathscr{G}$ and $\xi_k = \xi$.*

With that we define the ranked disjunction of two graphs $\mathscr{G}$ and $\mathscr{G}'$. The ranked disjunction is a way of merging two different priority graphs in a way that respects the degrees of the formulas of each graph. This will be used a step to merge the chains in $[\![\varphi]\!]$ and $[\![\neg\varphi]\!]$ when we perform the lexicographic contraction.

Definition 4.57 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$, $\mathscr{G}' = \langle \Phi', \prec' \rangle$ be a P-Graphs. The disjunction of $\mathscr{G}$ and $\mathscr{G}'$ is the P-Graph $\mathscr{G} \vee \mathscr{G}' = \langle \Phi \vee \Phi', \prec \vee \prec' \rangle$ where:*

$$\begin{aligned} \Phi \vee \Phi' &= \{\xi \vee \xi' \mid \xi \in \Phi \text{ and } \xi' \in \Phi'\} \\ \prec \vee \prec' &= \{\langle \xi_1 \vee \xi_1', \xi_2 \vee \xi_2' \rangle \in \Phi \vee \Phi' \mid d_{\mathscr{G}}(\xi_1) + d_{\mathscr{G}'}(\xi_1') < d_{\mathscr{G}}(\xi_2) + d_{\mathscr{G}'}(\xi_2')\} \end{aligned}$$

We will construct the lexicographic contraction of a graph $\mathscr{G}$ by a formula φ by merging the orders in the supports of the formulas φ and $\neg\varphi$. To do this, first we create a partially ordered (p.o.) set which corresponds to an intermediate graph in which the order relation is not strict, as in a P-Graph, and later we cluster the nodes in a same equivalence class to form a strict ordered graph, which will be the resulting P-graph.

Definition 4.58 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph in normal form and φ a propositional formula. The symmetric contraction of $\mathscr{G}$ by φ, is the p.o set $(\mathscr{G} \perp \varphi) = \langle \Phi \perp \varphi, \preceq' \rangle$ where:*

$$\begin{aligned} \Phi \perp \varphi &= \Phi_\varphi \vee \Phi_{\neg\varphi} \\ \xi_{\varphi_1} \vee \xi_{\neg\varphi_1} \preceq_{\perp\varphi} \xi_{\varphi_2} \vee \xi_{\neg\varphi_2} \quad \textit{iff} \quad & \xi_{\varphi_1} \vee \xi_{\neg\varphi_1} (\prec_\varphi \vee \prec_{\neg\varphi}) \xi_{\varphi_2} \vee \xi_{\neg\varphi_2} \textit{ or} \\ & (\xi_1 \prec \xi_2 \textit{ or } \xi_1' \prec' \xi_2') \textit{ and } d_{\mathscr{G}_\varphi}(\xi_{\varphi_1}) + d_{\mathscr{G}_{\neg\varphi}}(\xi_{\neg\varphi_1}) = \\ & d_{\mathscr{G}_\varphi}(\xi_{\varphi_2}) + d_{\mathscr{G}_{\neg\varphi}}(\xi_{\neg\varphi_2}) \end{aligned}$$

Finally, we construct the P-graph corresponding to the lexicographic contraction of a propositional formula φ from the graph $\mathscr{G}$ by joining all equivalent nodes in the symmetric contraction defined above, by means of disjunction. To make the definition more readable, we will use the notation $[\varphi]_\preceq = \{\xi \in \Phi \mid \xi \preceq \varphi \text{ and } \varphi \preceq \xi\}$ to denote the equivalence class of φ in the p.o. set $\langle \Phi, \preceq \rangle$.

Definition 4.59 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph whose propositions form a partition of the logical space and φ a propositional formula. We define the lexicographic contraction of φ from $\mathscr{G}$ as the P-Graph $\mathscr{G} \Downarrow \varphi = \langle \Phi', \prec' \rangle$ s.t.*

$$\Phi' = \{\bigvee [\xi]_{\preceq_{\perp\varphi}} \mid \xi \in \Phi \perp \varphi\}$$

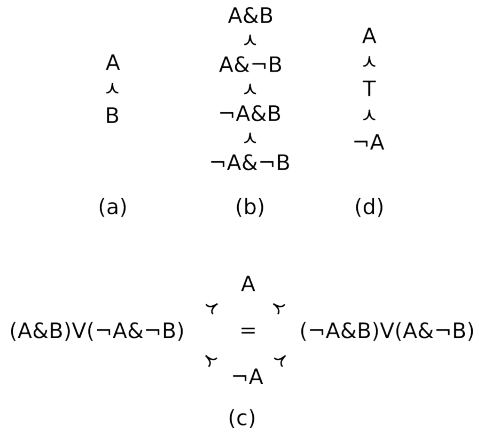
$$\bigvee [\xi]_{\preceq_{\perp\varphi}} \prec' \bigvee [\psi]_{\preceq_{\perp\varphi}} \qquad \textit{iff} \qquad \xi \preceq_{\perp\varphi} \psi \textit{ and } \psi \not\preceq_{\perp\varphi} \xi$$

To make it more concrete, we present the following example.

Example 4.60 *Lets take the graph $\mathscr{G}$ constituted by two nodes $A \prec B$. A model induced by such a graph is the model containing four worlds $A \wedge B < A \wedge \neg B < \neg A \wedge B < \neg A \wedge \neg B$. The*

lexicographic contraction of such model by the formula B, as defined in Section 4.1, would result in the model $A \wedge B \equiv A \wedge \neg B < \neg A \wedge B \equiv \neg A \wedge \neg B$. *The computation of the lexicographic contraction defined of B from the graph* $\mathscr{G}$ *above is depicted in the Figure 4.15*[10]*, where (a) represents graph* $\mathscr{G}$*; (b) the transformation of graph* $\mathscr{G}$ *into an equivalent graph* $\mathscr{G}'$ *partitioning the logical space; (c) the symmetric contraction of* $\mathscr{G}'$ *by formula B, and, (d) the resulting graph after lexicographic contraction.*

Figure 4.15 – Contracting graph $\mathscr{G}$ by formula B from Example 4.60.



Source: the author.

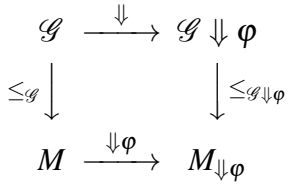
We point out that each element $\bigvee[\xi]_{\preceq_{\perp}\varphi}$ in the graph $\mathscr{G} \Downarrow \varphi$ corresponds to the notion "all the worlds with degree lesser then i", where $i = dg_{\mathscr{G}_\varphi}(\xi_\varphi) + dg_{\mathscr{G}_{\neg\varphi}}(\xi_{\neg\varphi})$ and $\xi = \xi_\varphi \vee \xi_{\neg\varphi}$. This is easy to see, since for any $\xi_\varphi \in \mathscr{G}_\varphi$, with $dg_{\mathscr{G}_\varphi}(\xi_\varphi) \leq i$, we can take a $\xi_\varphi \in \mathscr{G}_{\neg\varphi}$ such that $dg_{\mathscr{G}_{\neg\varphi}}(\xi_{\neg\varphi}) = i - dg_{\mathscr{G}_\varphi}(\xi_\varphi)$. As such, any world w in a model M induced by $\mathscr{G}$ with degree lesser then i satisfies one such element so that $M, w \vDash \bigvee[\xi]_{\preceq_{\perp}\varphi}$.

Fact 4.61 *Let* $\mathscr{G}$ *be a priority graph and* $M = \langle W, \leq, v\rangle$ *a preference model induced by* $\mathscr{G}$*. Let* $\xi = \xi_\varphi \vee \xi_{\neg\varphi}$ *be an element of* $\mathscr{G}_\varphi \vee \mathscr{G}_{\neg\varphi}$*, for some propositional formula* φ*. Any world* $w \in W$*, s.t.* $M, w \vDash dg_\varphi(i)$ *or* $M, w \vDash dg_\varphi(i)$*, with* $i \leq dg_{\mathscr{G}_\varphi}(\xi_\varphi) + dg_{\mathscr{G}_{\neg\varphi}}(\xi_{\neg\varphi})$*, then* $M, w \vDash \bigvee[\xi]_{\preceq_{\perp}\varphi}$ *as defined in Definition 4.59.*

We can prove harmony between the semantically defined lexicographic contraction on preference models and syntactic transformation in priority graphs. The intuition behind the proof is that the support of φ in a graph $\mathscr{G}$ will represent the chains in $[\![\varphi]\!]$ in the induced model. Similarly for the support of $\neg\varphi$. Taking the disjunction of both subgraphs, we conflate the chains in a way that if a world belongs in a chain of size i in either $[\![\varphi]\!]$ or $[\![\neg\varphi]\!]$, it will be

[10] In the figure, an arrow starting in node a and ending in node b represents the relation $a \prec b$ of the graph.

Figure 4.16 – Harmony for Lexicographic Contraction.



Source: the author

preferred to any world belonging in a chain of size $j > i$ in either $[\![\varphi]\!]$ or $[\![\neg\varphi]\!]$, as required by the definition of lexicographic contraction.

Theorem 4.62 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph in normal form, $M = \langle W, \leq, v \rangle$ a model induced by $\mathscr{G}$ and φ a propositional formula. If $\mathscr{G}_\varphi$ and $\mathscr{G}_{\neg\varphi}$ are adequate with respect to the M_φ and $M_{\neg\varphi}$, the model $M_{\Downarrow\varphi}$ is induced by the graph $\mathscr{G} \Downarrow \varphi$. In other words, the diagram in Figure 4.16 commutes.*

Proof: Immediate from Fact 4.61. □

With this operation, now we have four harmonic operations that can be used to specify changes in an agent's mental state: public announcement, or knowledge acquisition; radical upgrade, or preference adoption; public suggestion, or preference update; and lexicographic contraction, or preference contraction. These operations are enough to represent most of the changes in an agent mental state, such as belief update and contraction, desire addition and removal and intention adoption and dropping.

While we only presented axiomatizations for Preference Logic extended with each specific dynamic operation, we point out that a complete Dynamic Preference Logic consisted by extending Preference Logic $\mathscr{L}_{\leq}(P)$ with all the harmonic operations presented earlier can be achieved by simply joining the axiomatizations established in Propositions 4.25, 4.30, 4.35, 4.40, 4.44 and 4.50.

4.6 The perks of being a broad model

From Theorem 4.62, we know that to provide a codification of Lexicographic Contraction by means of transformation in P-graphs, the P-graphs and the induced models must have a strong relation, which we call the P-graph being adequate in respect to the model. In this section, we wish to investigate a class of well-founded preference models for which any P-graph inducing one of such models is adequate with respect to it. In other words, we wish to constrain the kind of models we work with to guarantee we can always perform contraction operations by transformations on P-graphs. These models are what we introduced in Section 4.3 as broad models.

Notice that the Theorem 4.41 (and similarly Moderate Contraction) is a very serious indication that contraction operations, in general, are very difficult - when not impossible - to characterize by means of operations on priority graphs. This is because, in a way, natural contraction is the "minimal operation" of iterated contraction one can define, in the sense that the only changes performed in the model are those required to guarantee that the for any model $M = \langle W, \leq, v \rangle$, the set $Min_{\leq_{\downarrow\varphi}} W = Min_{\leq} W \cup Min_{\leq} [\![\neg\varphi]\!]$, as required by AGM postulates (GROVE, 1988).

While pursuing a syntactic codification of lexicographic contraction, however, we have stumbled into an effective condition to provide such a codification - namely the adequateness of the support graphs. Notice that the root of the result in Theorem 4.41 is that we can always construct a model M_2 for which the priority graph $\mathscr{G}$ is not adequate and, as such, we cannot guarantee that a syntactic codification of natural contraction applied to $\mathscr{G}$ will properly describe the changes in M_2. If we restrict the analysis to only those models which are adequate, however, Lemma 4.54 tells us that such a codification exists, i.e. for adequate priority graphs Natural Contraction can be harmonic[11].

If we take the disjunction of the elements with syntactic depth 1 in the graph $\mathscr{G}_{\neg\varphi}$, by Lemma 4.54 it represents exactly the minimal elements satisfying $\neg\varphi$ of any induced model M for which $\mathscr{G}$ ($\mathscr{G}_{\varphi}$ and also $\mathscr{G}_{\neg\varphi}$) is adequate. As such it is not difficult to provide such an encoding as in Definition 4.63.

The intricacies of the construction in Definition 4.63 will not be discussed, since this definition is but an illustration of the possibility to define natural contraction. Nevertheless, the idea behind the construction in Definition 4.63 is that we will include the minimal elements of

[11] Notice this is not a contraction to our claim to provide a negative solution to the question posed by Liu (2011), since harmony only holds for some specific pairs of priority graphs and models, not in general

the support graph $\mathscr{G}_{\neg\varphi}$ as minimal elements in $\mathscr{G}$.

Definition 4.63 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph in normal form. We define the natural contraction of $\mathscr{G}$, denoted by $\mathscr{G} \downarrow \varphi = \langle \Phi_{\downarrow\varphi}, \prec' \rangle$ as:*

- $\Delta = \{\xi \in \Phi \mid d_{\mathscr{G}}(\xi) = 1\}$
- $\Gamma = \{\xi \in \Phi \mid d_{\mathscr{G}_{\neg\varphi}}(\xi \wedge \neg\varphi) = 1\}$
- $\Gamma \wedge \psi = \{\xi \wedge \psi \mid \xi \in \Gamma\}$
- $\Phi_{\downarrow\varphi} = (\Phi \setminus (\Delta \cup \Gamma)) \cup (\Gamma \wedge \varphi) \cup \{\bigvee(\Delta \cup (\Gamma \wedge \neg\varphi)\}$
- $\xi \prec' \psi = \begin{cases} \xi \prec \psi & \text{if } \xi, \psi \in \Phi \setminus (\Delta \cup \Gamma) \\ \xi \prec \psi' & \text{if } \xi \in \Phi \setminus (\Delta \cup \Gamma), \psi = \psi' \wedge \varphi \text{ and } \psi' \in \Gamma \\ \xi' \prec \psi & \text{if } \psi \in \Psi \setminus (\Delta \cup \Gamma), \xi = \xi' \wedge \varphi \text{ and } \xi' \in \Gamma \\ \xi = \bigvee(\Delta \cup (\Gamma \wedge \neg\varphi) & \text{otherwise} \end{cases}$

Giving the insight of Lemma 4.54, it is not difficult to prove the following result.

Theorem 4.64 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$ be a P-Graph in normal form, $M = \langle W, \leq, v \rangle$ a model induced by $\mathscr{G}$ and φ a propositional formula. If $\mathscr{G}_\varphi$ and $\mathscr{G}_{\neg\varphi}$ are adequate with respect to the M_φ and $M_{\neg\varphi}$, the model $M_{\downarrow\varphi}$ is induced by the graph $\mathscr{G} \downarrow \varphi$.*

We know, by Lemma 4.55 that, given a model, we can always get a P-graph satisfying the conditions of Theorems 4.62 and 4.64. We wish to establish a similar correspondence in the opposite direction, i.e. giving a P-graph, how to obtain an induced model that these conditions are satisfied.

In fact, if a P-graph $\mathscr{G}$ in normal form does not have propositionally valid nor unsatisfiable elements, we can always guarantee the existence of a model for which the conditions of Theorems 4.62 and 4.64 hold, namely broad models. First notice that the neither the condition of normal form and of not having propositionally valid nor unsatisfiable elements are restrictions on the P-graphs we are working on.

For normal form this property has been proved by Liu (2011) in Lemma 4.51. We now prove that removing propositionally valid and propositionally unsatisfiable formulas of a graph does not change the set of induced models.

Fact 4.65 *Let $\mathscr{G} = \langle \Phi, \prec \rangle$, $\mathscr{G}' = \langle \Phi', \prec' \rangle$ be P-graphs, s.t. $\Phi' = \{\xi \in \Phi \mid \nvDash_{\mathscr{L}_0} \xi \leftrightarrow \bot$ and $\nvDash_{\mathscr{L}_0} \xi \leftrightarrow \top\}$ and $\phi \prec' \psi$ only if $\psi \prec \psi$. Then a preference model $M = \langle W, \leq, v \rangle$ is induced by $\mathscr{G}$ iff it is induced by the P-graph $\mathscr{G}' = \langle \Phi', \prec \rangle$.*

Proof: Notice that for any $\xi \in \Phi$ such that $\models_{\mathscr{L}_0} \xi \leftrightarrow \top$, i.e. ξ is propositionally valid, than for any $w \in W$, $M, w \models \xi$. Also, for any $\xi \in \Phi$ such that $\models_{\mathscr{L}_0} \xi \leftrightarrow \bot$, i.e. ξ is propositionally unsatisfiable, than there is no $w \in W$, s.t. $M, w \models \xi$. From these observations, the proof follows from the application of Definition 4.16 for induced model. □

Now, we introduced broad models in Definition 4.20 in Section 4.3 with the motivation of allowing to check the satisfiability of conditional preferences by means of propositional satisfiability checks. There are, however, other advantages in working with these models. The most important for us is the fact that for any priority graph $\mathscr{G}$ in normal form that does not have propositionally valid nor unsatisfiable elements that induces a broad model M, and any propositional formula φ, $\mathscr{G}$ is appropriate in regards to M, as are $\mathscr{G}_\varphi$ and $\mathscr{G}_{\neg\varphi}$ in regard to the restrictions M_φ and $M_{\neg\varphi}$.

Fact 4.66 *Let $\mathscr{G}$ be a priority graph in normal form that does not have propositionally valid nor unsatisfiable elements, $M = \langle W, \leq, v\rangle$ a broad preference model induced by $\mathscr{G}$ and $\varphi \in \mathscr{L}_0$ a propositional formula. $\mathscr{G}$ is appropriate in regards to M, $\mathscr{G}_\varphi$ is appropriate in regards to the restriction of M to the worlds satisfying φ and $\mathscr{G}_{\neg\varphi}$ is appropriate in regards to the restriction of M to the worlds not satisfying φ.*

Proof: This is a simple result, since all possible propositional valuations are represented in M, so if $\xi \in \mathscr{G}$, there is a world w in M, s.t. $M, w \models \xi$. Regarding the support graph $\mathscr{G}_\varphi$ (similarly $\mathscr{G}_{\neg\varphi}$), for any $\xi \in \mathscr{G}$ s.t. $\xi \nvDash_{\mathscr{L}0} \neg\varphi$ ($\xi \nvDash_{\mathscr{L}0} \varphi$), it means that $\xi \wedge \varphi$ ($\xi \wedge \neg\varphi$) and there is a world in M satisfying it, as such $\mathscr{G}_\varphi$ ($\mathscr{G}_{\neg\varphi}$) is appropriate. □

Fact 4.66 above states that, if we concern ourselves only broad models, we can perform contractions by means of operations in the priority graphs and guarantee its equivalence to the operations performed in the preference models. This is another indication that broad models are ideal representations if we wish to use priority graphs as data structures to implement our programming language in Chapter 6.

In the next chapter we will show how to use Dynamic Preference Logic as a language to specify agents. For that, we will propose a logic with two preference modalities representing the agents beliefs and desires.

4.7 Summary of the chapter

In this chapter, we present the language and semantics of Dynamic Preference Logic, a dynamic modal propositional logic following the dynamic Epistemic Logic tradition. DPL has already been employed to study several mental attitudes in the literature. Particularly important in this exposition is the connection between the preference models used to define the semantics of that logic and the syntactic structure of priority graphs, as defined by Liu (2011).

In Section 4.1, we present the basic language of Preference Logic with the further restriction on the models to satisfy the well-known Lewis' Limit Assumption, and provide a complete axiomatization for that logic. In Section 4.4, we extend this language with dynamic modalities already proposed in the literature representing different update operators representing well-know epistemic change operations. In Section 4.5, we propose contraction operators for Dynamic Preference Logic and show that some of them cannot be defined by means of priority graphs. This result is a general answer to the question posed by Liu (2011) about whether any PDL-definable operation preserving preference models is harmonic, i.e. can be equivalently represented by transformations in priority graphs.

In Table 4.1, we present the operations studied in this chapter as well as their properties regarding the possibility of defining the operation by means of a PDL program that holds for any preference model or just for broad models and the harmony properties, i.e. if they can be defined by means of transformations on priority graphs, in regards to any preference model or just to broad models.

Table 4.1 – Operations studied in this Chapter and their properties

Operation	PDL-definable over all preference models	PDL-definable over broad models	Harmonic for all preference models	Harmonic for broad models
Public Announcement	Yes	Yes	Yes	Yes
Radical Upgrade	Yes	Yes	Yes	Yes
Public Suggestion	Yes	Yes	Yes	Yes
Natural contraction	Yes	Yes	No	Yes
Moderate contraction	Yes	Yes	No	Yes
Lexicographic contraction	No	Yes	Unknown	Yes

Source: the authors

5 A LOGIC FOR THE DYNAMICS OF MENTAL ATTITUDES

Once established the language and semantics of Dynamic Preference Logic, we will now use this language to model the mental attitudes we are interested in. Thus, in this chapter we propose a logic for representing mental attitudes and their dynamics for Agent Programming, based on the formalism of Dynamic Preference Logic. As done for the presentation of Dynamic Preference Logic, in Chapter 4, we will first introduce the static part of the language and study the representation of mental attitudes in this logic and, later, we will introduce change operations to our logic.

In Section 5.1, this new logic is constructed by using Preference Logic as a language to represent the agent's epistemic and conative state, i.e. their mental state concerning beliefs and desires about the world. We will model the epistemic and conative state of an agent by means of preference relations of plausibility and desirability, respectively, as also proposed by Boutilier (1994b). In this logic, we show how to encode the notions of '*it is known that* φ', '*it is believed that* φ' and '*it is desired that* φ' as formulas $K\varphi$, $B\varphi$ and $G\varphi$[1], respectively. In Section 5.2, we analyse an encoding of the notion of intention in the logic and, giving the requirements for a formalization of intention discussed in Chapters 2 and 3, we also propose an extension of our logic to represent agent's plans.

Finally, in Section 5.3, this logic will be dynamified, in a similar way to what has been done in the previous chapter, resulting in a logic for the dynamics of the mental attitudes of agents.

5.1 A logic of beliefs and desires

As discussed in Chapter 4, Preference Logic has been used to encode several different mental attitudes in the literature, among them knowledge, beliefs (BALTAG; SMETS, 2008) and goals or desires (BOUTILIER, 1994b; LANG; VAN DER TORRE; WEYDERT, 2003). In this chapter, we propose a logic encoding both notions simultaneously, similar to the logic proposed by Boutilier (1994b) and by Liu (2011).

For that, we will need two preference orderings in the models: one for encoding the notion of *plausibility* or *doxastic normality*, written $\leq_P$, which will be used to encode beliefs,

[1] As is common practice in Agent Programming, we will refer to the agent desires as 'goals.' The exact relation between the notions of 'goal' and 'desire' varies in the literature, but in our work they will stand for the motivational attitude in the agent's mind. We discuss this further in Section 5.1

and one for *preference* or *desirability*, written $\leq_D$, which will be used to encode desires. Once established this support language, constituted by the combination of the individual Preference Logics for plausibility and desirability, we will provide encodings for mental attitudes.

Definition 5.1 *We define the language $\mathscr{L}_{\leq_P,\leq_D}(P)$ by the following grammar (where $p \in P$ a set of propositional letters):*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid A\varphi \mid [\leq_P]\varphi \mid [<_P]\varphi \mid [\leq_D]\varphi \mid [<_D]\varphi$$

As before, we will define $E\varphi \equiv \neg A\neg\varphi$ and $\langle\leq\rangle\varphi \equiv \neg[\leq]\neg\varphi$. The formula $[\leq_D]\varphi$ ($[\leq_P]\varphi$) means that in all words equally or more desirable (plausible) than the current one, φ holds and $[<_D]\varphi$ ($[<_P]\varphi$) that in all words strictly more desirable (plausible) than the current one, φ holds.

To interpret these formulas, we will introduce a new kind of Kripke model containing two accessibility relations - one for plausibility and one for desirability. We will call this new model an *agent model*.

Definition 5.2 *An agent model is a tuple $M = \langle W, \leq_P, \leq_D, v\rangle$ where W is a set of possible worlds, and both $\leq_D$ and $\leq_P$ are pre-orders over W with well-founded strict parts.*

Notice that an agent model is an amalgamation of two different preference models encoding the orderings for plausibility and desirability. As such, the relation $\leq_P$ represents a plausibility relation between worlds describing which worlds are more plausible to be the actual state of affairs. This relation encode the doxastic state of the agent, i.e. her beliefs about the world. As such, $w \leq_P w'$ means that it is more plausible for the actual world to be w than it is for the actual world to be w'. Similarly, the relation $\leq_D$ describes the notion of desirability, meaning which worlds the agent would prefer to be actual world, encoding then the notion of desire. As such, $w \leq_D w'$ means that it is more desirable for the actual world to be w than it is for the actual world to be w'.

The interpretation of the formulas is defined as usual. We will only present the interpretations for the modalities, since the semantics of the propositional connectives is clear. They are interpreted as:

$$\begin{aligned}
M, w \vDash [\leq_P]\varphi \quad &\text{iff } \forall w' \in W : w' \leq_P w \Rightarrow M, w' \vDash \varphi \\
M, w \vDash [<_P]\varphi \quad &\text{iff } \forall w' \in W : w' <_P w \Rightarrow M, w' \vDash \varphi \\
M, w \vDash [\leq_D]\varphi \quad &\text{iff } \forall w' \in W : w' \leq_D w \Rightarrow M, w' \vDash \varphi \\
M, w \vDash [<_D]\varphi \quad &\text{iff } \forall w' \in W : w' <_D w \Rightarrow M, w' \vDash \varphi
\end{aligned}$$

An axiomatization for the logic is provided in Figure 5.1. This axiomatization is achieved by replicating the axiomatization presented in Figure 4.1 for the static Preference Logic in Chapter 4 for both the modalities $[\leq_P]$ and $[\leq_D]$. The completeness of this axiomatization is guaranteed by the following transfer result in fusion of modal logics.

Theorem 5.3 *(BLACKBURN; VAN BENTHEM; WOLTER, 2006) If the modal logics L_1 and L_2 are characterised by classes of frames $\mathscr{C}_1$ and and $\mathscr{C}_2$, respectively, and if $\mathscr{C}_1$ and and $\mathscr{C}_2$ are closed under the formation of disjoint unions and isomorphic copies, then the fusion $L_1 \otimes L_2$ is characterized by the class of models*

$$\mathscr{C}_1 \otimes \mathscr{C}_2 = \{\langle W, R_1, \ldots, R_n, S_1, \ldots, S_m\rangle \mid \langle W, R_1, \ldots, R_n\rangle \in \mathscr{C}_1 \text{ and } \langle W, S_1, \ldots, S_m\rangle \in \mathscr{C}_2\}$$

As the axiomatization of the fusion $L_1 \otimes L_2$ is the union of the axiomatizations of L_1 and L_2, the following result ensues.

Lemma 5.4 *The axiomatization presented in Figure 5.1, taken together with the propositional validities and the Modus Ponens and the Necessitation rules, is sound and complete for the class of agent models.*

Proof: The axiomatization presented in Figure 5.1 is the union of the axiomatizations for Preference Logic presented in Figure 4.1 in Chapter 4 for each modality. Also, given an agent model $M = \langle W, \leq_P, \leq_D, v\rangle$, clearly $\langle W, \leq_P, v\rangle$ is a preference model for modality $[\leq_P]$ and $\langle W, \leq_D, v\rangle$ is a preference model for modality $[\leq_D]$. As such, by Theorem 5.3, the axiomatization depicted in Figure 5.1 expanded with all propositional validities and the necessitation and *modus ponens* rules is a sound and complete axiomatization of $\mathscr{L}_{\leq_P, \leq_D}(P)$. □

Now that we have established the logic we will work with, let's proceed with the representation of an agent model by means of syntactical structures, as it has been done in Chapter 4 for preference models. These structures will be used to implement the semantics an agent programming language in Chapter 6 and serve as basis to reason about the agent's mental state in that semantics.

From Chapter 4, we know that preferences models have a syntactical counterpart in priority graphs. Since agent models are nothing more that the union of two preference models, we know that there must be a similar syntactic representation for agent models as well. We will define, thus, the notion of an agent structure, which will serve as this syntactic counterpart for agent models.

Figure 5.1 – Axiomatization of the logic of plausibility and desirability.

$$
\begin{array}{ll}
\mathbf{K}_P: & [\leq_P](\varphi \to \psi) \to ([\leq]\varphi \to [\leq]\psi) \\
\mathbf{T}_P: & [\leq_P]\varphi \to \varphi \\
\mathbf{4}_P: & [\leq_P]\varphi \to [\leq_P][\leq_P]\varphi \\
 & \\
\mathbf{K}_{N^<}: & [<_P](\varphi \to \psi) \to ([<_P]\varphi \to [<_P]\psi) \\
\mathbf{W}_{N^<}: & [<_P]([<_P]\varphi \to \varphi) \to [<_P]\varphi \\
PP^<{}_1: & [\leq_P]\varphi \to [<_P]\varphi \\
PP^<{}_2: & [<_P]\varphi \to [<_P][\leq_P]\varphi \\
PP^<{}_3: & [<_P]\varphi \to [\leq_P][<_P]\varphi \\
PP^<{}_4: & [\leq_P]([\leq_P]\varphi \vee \psi) \wedge [<_P]\psi \to \varphi \vee [\leq_P]\psi \\
 & \\
\mathbf{K}_D: & [\leq_D](\varphi \to \psi) \to ([\leq_D]\varphi \to [\leq_D]\psi) \\
\mathbf{T}_D: & [\leq_D]\varphi \to \varphi \\
\mathbf{4}_D: & [\leq_D]\varphi \to [\leq_D][\leq_D]\varphi \\
 & \\
\mathbf{K}_{P^<}: & [<_D](\varphi \to \psi) \to ([<_D]\varphi \to [<_D]\psi) \\
\mathbf{W}_{P^<}: & [<_D]([<_D]\varphi \to \varphi) \to [<_D]\varphi \\
DD^<{}_1: & [\leq_D]\varphi \to [<_D]\varphi \\
DD^<{}_2: & [<_D]\varphi \to [<_D][\leq_D]\varphi \\
DD^<{}_3: & [<_D]\varphi \to [\leq_D][<_D]\varphi \\
DD^<{}_4: & [\leq_D]([\leq_D]\varphi \vee \psi) \wedge [<_D]\psi \to \varphi \vee [\leq_D]\psi \\
 & \\
\mathbf{K}_A: & A(\varphi \to \psi) \to (A\varphi \to A\psi) \\
\mathbf{T}_A: & A\varphi \to \varphi \\
\mathbf{4}_A: & A\varphi \to AA\varphi \\
\mathbf{B}_A: & \varphi \to A\neg A\neg\varphi \\
AP: & A\varphi \to [\leq_P]\varphi \\
AD: & A\varphi \to [\leq_D]\varphi
\end{array}
$$

Source: the author.

Definition 5.5 *Let $\mathscr{L}_0(P)$ be the propositional language constructed over the set of propositional letters P, as usual. An agent structure is a pair $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D \rangle$, where both $\mathscr{G}_P = \langle \Phi_P, \prec_P \rangle$ and $\mathscr{G}_D = \langle \Phi_D, \prec_D \rangle$ are P-graphs.*

From agent structures we define the notion of induced agent model, similar to what was done to preference models in Definition 4.16. We just need to take the P-graphs that induce the plausibility and desirability relations ($\leq_P$ and $\leq_D$, respectively) which are guaranteed to exist by Theorem 4.17.

Definition 5.6 *Let $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D \rangle$ be an agent structure and $M = \langle W, \leq_P, \leq_D, v \rangle$ an agent model. We say M is induced by $\mathscr{G}$ iff $\leq_P = \leq_{\mathscr{G}_P}$ and $\leq_D = \leq_{\mathscr{G}_D}$.*

From Definition 5.6, it is clear that every agent model is induced by some agent structure.

Theorem 5.7 *Let $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model. There is an agent structure $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D\rangle$ s.t. M is induced by $\mathscr{G}$.*

Proof: The result follows immediately from the fact that both $\langle W, \leq_P, v\rangle$ and $\langle W, \leq_D, v\rangle$ are preference models and from Theorem 4.17. □

In the following we will use the language $\mathscr{L}_{\leq_P, \leq_D}$ to encode the notions of knowledge, belief and goal (or desire) we will adopt in our work.

5.1.1 Encoding knowledge, belief and goal

In this section, we aim to encode in the language $\mathscr{L}_{\leq_P, \leq_D}$ the mental attitudes of knowledge, belief and goal as commonly used in Agent Programming. We will use conditional modellings of the mental attitudes represented by dyadic modalities, except for the notion of knowledge.

As discussed in Section 4.1, conditional modalities as $C(\psi|\varphi)$ are claimed to be more faithful representation of mental attitudes, as held by humans. Also, these conditional modalities are natural ways to encode non-monotonic reasoning, such as non-monotonic inference rules $A \Rightarrow B$ widely used in agent programming languages, e.g. to enrich the agent's belief base (BORDINI; HUBNER; WOOLDRIDGE, 2007), and to specify goal selection rules (DASTANI et al., 2003).

We wish to point out that, in this work, we assume no conceptual difference between the terms 'desire' and 'goal'. Some authors distinguish the notion of goals from that of desires, claiming that goals are a consistent subset of agent's desires (RAO; GEORGEFF, 1998). This differentiation is mainly due to the fact that agent's desires need not to be consistent, which poses difficulties to monotonic logical representations of these attitudes, such as proposed by Rao and Georgeff (1998). As our encoding of mental attitudes is non-monotonic, we can represent inconsistent desires without the risk of trivializing the agents desires.

As the term 'goal' is widely used in Agent Programming to represent an agent's motivational attitudes, we will adopt the terminology 'goal' to stand for the motivational attitude of the agent, i.e. for desires in BDI theory. This terminological choice has also the additional advantage of differentiating 'desires' and 'desirability', in the sense that what is desired is not necessarily what is always most desirable, as we will see further.

Let's start with encoding knowledge. In our work, the notion of knowledge is equal to that of (global) epistemic necessity. As such, we will say that '*it is known that* 'φ by the

formula $A\varphi$. As mentioned in Chapter 4, epistemic necessity is represented in our logic by the modality A, i.e. in our logic $A\varphi$ means that in all conceivable worlds to the agent, the formula φ is true. In this sense, the notion of knowledge adopted here is a strengthening of the notion of knowledge as a unrevisable true belief (or safe belief) proposed by Baltag and Smets (2008).

Let's examine the more interesting case of beliefs. We want to define a conditional modality $B(\psi|\varphi)$, similar to $C(\psi|\varphi)$ in Chapter 4, meaning '*in the most plausible* φ *worlds*, ψ *holds*.' We propose the following codification of conditional belief:

$$B(\psi|\varphi) \equiv A((\varphi \wedge \neg\langle\leq_P\rangle\varphi) \rightarrow \psi) \tag{5.1}$$

Notice that in the formula 5.1 above, the subformula $(\varphi \wedge \neg\langle\leq_P\rangle\varphi)$ has the same structure as the formula $\mu\varphi$ in Definition 4.3, using the modality $\leq_P$. Since we have two accessibility relations now, we will use $\mu_P\varphi$ to refer to the minimal worlds satisfying φ according to the relation $\leq_P$, and similarly $\mu_D\varphi$ to the relation $\leq_D$. As such, we can rewrite the formula 5.1 as:

$$B(\psi|\varphi) \equiv A((\mu_P\varphi) \rightarrow \psi) \tag{5.2}$$

Clearly, the semantics of $B(\psi|\varphi)$ implies that the most plausible φ-worlds are ψ-worlds, i.e. $Min_{\leq_P}[\![\varphi]\!] \subseteq [\![\psi]\!]$. Finally, we define the unconditional belief $B(\psi)$, meaning '*it is more plausible that* ψ *holds*', as:

$$B(\psi) \equiv B(\psi|\top) \tag{5.3}$$

Encodings of the notion of desire and goal are numerous in the literature with various meanings according to the intended application. For the purpose of Agent Programming, two non-monotonic representations of goals strike us as the most interesting: Boutilier (1994b)'s ideals and Van Riemsdijk, Dastani and Meyer (2009)'s goals. We will encode both proposals in our logic and they will both be of use in our work to model different motivational phenomena, namely the notions of 'desire' and 'overwhelming desire' from Philosophy of Action (c.f. Chapter 2).

Let's start with ideals. Boutilier (1994b) proposes the conditional modal $I(\psi|\varphi)$ meaning '*if* φ, *then ideally* ψ', to encode the notion of desire (or goal) in QDT. The meaning behind this statement is that in the most desirable φ worlds, ψ holds, much like the already proposed encoding for beliefs. Using a logic very similar to ours, Boutilier (1994b) encodes ideals by means of most-desirable worlds, which translated to our language may be encoded as:

$$I(\psi|\varphi) \equiv A(\mu_D\varphi \rightarrow \psi) \tag{5.4}$$

It is our belief that Boutilier's ideals model quite faithfully the notion of *overwhelming desire*, i.e. a desire that is always preferred to its alternatives. As such, the formula $I(\psi) \equiv I(\psi|\top)$ models the fact that the agent '*necessarily wants that* ψ', i.e. in the most desirable worlds ψ holds.

On the other hand, Van Riemsdijk, Dastani and Meyer (2009) propose a non-monotonic semantics for goals in Agent Programming which we believe to be a faithful encoding of motivational attitudes in the BDI paradigm. Let's review their semantics. An agent in their formalism is a pair of $\langle \sigma, \gamma \rangle$ of sets propositional formulas, where σ is the belief base of the agent and γ her goal base. As such, we say the agent has the goal to φ, denoted $\langle \sigma, \gamma \rangle \vDash G\varphi$ if we can derive φ from a consistent subset of her goal base γ, i.e.

$$\langle \sigma, \gamma \rangle \vDash G\varphi \quad \text{iff} \quad \exists \gamma' \subseteq \gamma \text{ s.t. } \gamma' \nvDash \bot \text{ and } \gamma' \vDash \varphi$$

To understand their semantics in our framework, however, we must provide a faithful interpretation of a given agent $\langle \sigma, \gamma \rangle$ as a agent model (or agent structure). While the belief base σ can be easily understood as the singleton priority graph $\mathscr{G}_\sigma = \langle \{\bigwedge \sigma\}, \emptyset \rangle$[2], since γ may be inconsistent this transformation does encode the same information as γ in Van Riemsdijk, Dastani and Meyer (2009)'s semantics. We propose that the goal base γ can be understood as the flat priority graph $\mathscr{G}_\gamma = \langle \gamma, \emptyset \rangle$.

Notice that, given a preference model $M = \langle W, \leq, v \rangle$ induced by the P-graph $\mathscr{G}_\gamma$, if there is a consistent subset γ' of γ, then there is a maximal subset γ'' of γ containing all formulas in γ'. If there is a world $w \in W$ satisfying all formulas of γ'', then w is minimal in M - given the structure of $\mathscr{G}_\gamma$ and Definition 4.14 of a preference relation induced by a graph. As such, the notion $\langle \sigma, \gamma \rangle \vDash G\varphi$, can be faithfully enconded by the notion 'there is a minimal world satisfying φ' in our semantics.

We propose, thus, the following codification of a conditional variant for goals in our preference logic, meaning that '*in the context of* φ *it is conceivably desirable to* ψ' or '*if* φ*, one can pursue the goal to* ψ':

$$G(\psi|\varphi) \equiv E(\mu_D \varphi \wedge \psi) \tag{5.5}$$

We claim that the goal modality G have similar characteristics to that notion of goal by Van Riemsdijk, Dastani and Meyer (2009). For example, both modalities satisfy exactly the same modal properties (VAN RIEMSDIJK; DASTANI; MEYER, 2009) of Chellas' classifica-

[2] As σ is consistent, the graph $\mathscr{G}_\sigma$ induces models whose minimal elements satisfy all formulas entailed by σ.

tion (CHELLAS, 1980).

Proposition 5.8 *The goal modality G defined in the formula 5.5 satisfies the axiom*

$$\mathbf{M}: G(\varphi \wedge \psi|\xi) \rightarrow (G(\varphi|\xi) \wedge G(\psi|\xi))$$

but not

$$\begin{aligned} \mathbf{K} &: G(\xi \rightarrow \psi|\varphi) \rightarrow (G(\xi|\varphi) \rightarrow G(\psi|\varphi)) \\ \mathbf{D} &: \neg(G(\xi|\varphi) \wedge G(\neg\xi|\varphi)) \\ \mathbf{C} &: (G(\xi|\varphi) \wedge G(\psi|\varphi)) \rightarrow G(\xi \wedge \psi|\varphi) \end{aligned}$$

We point out, however, that our encoding for goal is much more permissive than that of Van Riemsdijk, Dastani and Meyer (2009), in the sense that we allow much more formulas to be desired. In fact, for any formula φ we have that $G(\varphi) \vee G(\neg\varphi)$ is valid in $\mathscr{L}_{\leq_P,\leq_D}$ - a validity that does not hold in the semantics of Van Riemsdijk, Dastani and Meyer (2009)[3].

We believe this is not a semantic problem to our notion of goals, since the validity $G(\varphi) \vee G(\neg\varphi)$ only expresses that for any formula φ either the agent envision it as more desirable than $\neg\varphi$ (if $G(\varphi)$ and not $G(\neg\varphi)$), less desirable than $\neg\varphi$ (if $G(\neg\varphi)$ and not $G(\varphi)$), or indifferent of whether φ holds (if both $G(\varphi)$ and $G(\neg\varphi)$).

In the remainder of our work, we adopt formula 5.5 as the codification for the notion of goal/desire, and formula 5.4 as the codification for the notion of overwhelming desire. Notice that the modalities G and I are duals, in the sense that $G(\psi|\varphi) \Leftrightarrow \neg I(\neg\psi|\varphi)$. This means that the formula $G(\psi|\varphi)$ has a connotation that 'in the context of φ, it is possible that ψ is better' or, equivalently, 'in the context of φ, $\neg\psi$ is not necessarily better.'

Now that we have codifications for mental attitudes in the logic $\mathscr{L}_{\leq_P,\leq_D}$, we need to show how to reason about agents beliefs and goals by means agent structures, similarly as it has been done for conditional preferences and priority graphs in Proposition 4.19. We point out that the formula $\mu_{\langle\varphi_1,\ldots,\varphi_P\rangle}$ corresponds to the formula 4.1 defined in Chapter 4, representing the syntactic encoding of $\mu\varphi$ based on a priority graph.

Proposition 5.9 *Let $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model induced by the agent structure $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D\rangle$ and $w \in W$. Let yet, $\Pi_{\mathscr{G}_\square}$ denote the set of maximal paths $\langle\varphi_1,\ldots,\varphi_P\rangle$ in $\mathscr{G}_\square$ on P-graph $\mathscr{G}_\square$ (where $\square$ is either P or D). Then*

1. *$M, w \vDash B(\psi|\varphi)$ iff $M, w \vDash A((\bigvee_{\langle\varphi_1,\ldots,\varphi_n\rangle \in \Pi_{\mathscr{G}_P}} \mu_{\langle\varphi_1,\ldots,\varphi_n\rangle}\varphi) \rightarrow \psi)$*

[3]It is easy to provide an example that it doesn't hold Van Riemsdijk, Dastani and Meyer (2009)'s semantics. For example, take $\gamma = \{p\}$, for any σ, neither $\langle\sigma,\gamma\rangle \vDash G(q)$, nor $\langle\sigma,\gamma\rangle \vDash G(\neg q)$ hold.

2. $M, w \vDash G(\psi|\varphi)$ *iff* $M, w \vDash E((\bigvee_{\langle\varphi_1,\ldots,\varphi_n\rangle\in\Pi_{\mathscr{G}_D}} \mu_{\langle\varphi_1,\ldots,\varphi_n\rangle}\varphi) \wedge \psi)$
3. $M, w \vDash I(\psi|\varphi)$ *iff* $M, w \vDash A((\bigvee_{\langle\varphi_1,\ldots,\varphi_n\rangle\in\Pi_{\mathscr{G}_D}} \mu_{\langle\varphi_1,\ldots,\varphi_n\rangle}\varphi) \rightarrow \psi)$

Proof: The proof follow from the codification of $B(\varphi|\psi)$, $G(\varphi|\psi)$ and $I(\varphi|\psi)$, given by formulas 5.2, 5.5 and 5.4 respectively, and from Proposition 4.19. On the proof of the claim 2, we point out that $G(\psi|\varphi) \leftrightarrow \neg I(\neg\psi|\varphi)$. □

The theorem above guarantees that all the information about the minimal worlds can be described by means of the agent structure. As such, if we take as broad models those satisfying $W = 2^P$ and $v(p) = \{w \in W \mid p \in w\}$, the formulas above can be decided by propositional satisfiability and validity. This is because in these broad models, $M, w \vDash E\varphi$ iff φ is propositionally satisfiable and $M, w \vDash A\varphi$ if φ is valid. As such, we don't need a model to reason about the agent's beliefs or desires. All relevant information about the agent's mental state is encoded in the agent structure.

Definition 5.10 *Let $\mathscr{G}$ be an agent structure and $M = \langle W, \leq_P, \leq_D, v\rangle$ an agent model induced by $\mathscr{G}$. We say M is the broad model of $\mathscr{G}$ iff $W = 2^P$ and for all $p \in P$, $v(p) = \{w \in W \mid p \in w\}$.*

We will say an agent structure $\mathscr{G}$ proves a formula φ, denoted by $\mathscr{G} \vDash \varphi$, if the broad model of $\mathscr{G}$ proves φ, i.e. $M \vDash \varphi$.

Now we proceed to investigate a codification for intentions in this logic. For that, we will need to include the notion of actions (or plans) to represent the practical aspect of intentional action.

5.2 Introducing intentions in the logic of rationality

It is not yet consensus which properties a theory of intentions should satisfy to properly describe the notions of intentional action, intentionality, etc. In the Artificial Intelligence research, Cohen and Levesque (1990)'s desiderata for intentions based in Bratman (1999)'s work has become the official benchmark for any theory aiming to formalize such notions. As such, we will take these requirements as a guiding light to encode our notion of intention. Let's see them one more time.

1. Intentions normally pose problems for the agent, i.e. the agent needs to determine a way to achieve her intention;
2. Intentions provide a "screen of admissibility" for adopting other intentions;

3. Agents "track" the success of their attempts to achieve their intentions;
4. The agent believes her intentions are possible;
5. The agent does not believe she will not bring about her intentions;
6. Under certain conditions, the agent believes she will bring about her intention;
7. Agents need not intend all the expected side-effects of their intentions.

Central to these seven requirements, in our understanding, are two distinctive roles of intention in practical reasoning: the role of intention as a constraint in the possible actions/goals entertained by the agent (as requirements 2, 4 and 7) and intention as a product of practicality, i.e. intentions as intrinsically connected to plans (as in requirements 1, 3, 4, 5 and 6).

In this section, our main goal is to provide a conditional formula $Int(\psi|\varphi)$ meaning '*in the context of* φ, *the agent intends to* ψ'. First, since the logic of beliefs and goals developed in the previous section does not provide means to reason about practicality, i.e. about action and plans, we start by characterizing intention as a "window of acceptability' for new intentions and goals. Later, in Subsection 5.2.2, we extend the logic proposed above including ontic actions to express the relation between intention and practicality. Notice that following the BDI paradigm, this extension is a necessary one since intention is a *sui generis* mental attitude that cannot be reduced to the notions of goal and belief.

5.2.1 Intention as a "window of acceptability"

While we use Cohen and Levesque (1990)'s desiderata to guide the construction of our encoding of intention, we point out that these requirements enumerate the main points of the concept of intention in the work of Bratman (1999). As such, it is through the concept of intentions in Bratman (1999) that we propose and evaluate this encoding.

An intention, according to Bratman (1999), imposes a "window of acceptability" for other intentions, meaning that an intention affects the agent mental state in a way that other desires incompatible with them are not considered by the agent in the process of practical reasoning. According to Bratman, this is one fundamental role of intentions to allow rationality in resource bounded agents.

As discussed in Chapter 2, Bratman's strong consistency for intentions requires both internal consistency of the intentions and external consistency considering other intentions. Internal consistency means roughly that an intention is consistent in itself, i.e. the desired state of affairs is possible and the plans adopted to achieve it are consistent. Externally consistent, for

Bratman, means that the intentions held by the agent *taken together* are consistent.

From these requirements, we can obtain the following properties a formula $Int(\psi)$ must satisfy. By internal consistency, an intention is consistent within itself, i.e. the agent may not intend to achieve paradoxical states of affairs (or impossible worlds). This property can be represented in our language as

$$\neg Int(\bot)$$

By Vratman's external consistency, intentions are mutually consistent, thus an agent may not intend two mutually inconsistent states of affairs, i.e.

$$Int(\psi) \Rightarrow \neg Int(\neg\psi)$$

In regards to the connection between intentions and goals, as discussed in Chapter 2, we adopt the view that intending a state of affairs imply desiring it. In formal terms,

$$Int(\psi) \Rightarrow G(\psi)$$

Also, regarding the relationship between intentions and beliefs, the problem of modelling the properties expressed in the requirements above is a tricky one in the formalism we are using. Cohen and Levesque (1990)'s requirements 4, 5 and 6 presuppose that the agent refer to future states of affairs. In other words, intentions are necessarily towards the future, meaning that they represent states of affairs the agent desires and is committed to achieve. Since we don't have a temporal structure in our models, we cannot represent the agent's beliefs about the future.

Since requirements 5 and 6 of the Cohen and Levesque (1990)'s desiderata can only be represented with such a temporal structure or with the means to express agents actions, we will only tackle the requirement 4. We can represent a simplified form of belief-intention consistency given by epistemic possibility.

$$Int(\psi) \Rightarrow E\psi$$

The formula above represents that if an agent intends that ψ, then the agent believes ψ to be possible, i.e. an epistemically possible[4] state of affairs.

A more adequate solution satisfying requirements 5 and 6 will be proposed in Subsection 5.2.2 with the introduction of practicality in the language, since we will be able to represent

[4]Remember, $E\psi \equiv \neg A\neg\psi$, meaning that $\neg\psi$ is not epistemically necessary, thus ψ is epistemically possible.

that the agent believes the intended state of affairs to be achievable by the execution of some plan.

Well, an immediate candidate for an encoding of the formula $Int(\psi)$ comes to mind by making the identification of intentions as epistemically possible overwhelming desires, i.e. $Int(\psi) \equiv I(\psi) \wedge E\psi$. Notice that in our logic the following hold.

$$I(\psi|\varphi) \wedge E(\varphi \wedge \psi) \Rightarrow G(\psi|\varphi) \qquad \text{and} \qquad I(\psi|\varphi) \wedge E(\varphi \wedge \psi) \Rightarrow \neg I(\neg\psi|\varphi).$$

The first means that if in context of φ, the agent ideally prefers ψ (has an overwhelming desire to ψ), and φ and ψ are jointly possible, then she will actually desire (have a goal so that) ψ given φ. Additionally, the second states that if in context of φ the agent ideally prefers ψ and φ and ψ are jointly possible, then in this context, she does not ideally prefers $\neg\psi$.

One may argue that this proposal is directly incompatible with the supporting philosophical concept, given Bratman's critique of intentions as overwhelming desires. We point out that his critique concerns the desirability status of the alternatives before committing to an intention. In here, we are trying to encode the choices that the agent has already made, i.e. the intentions she is committed with. Since we can see committing to an intention as a mental state changing action (MEYER; VAN DER HOEK; VAN LINDER, 1999), it is not unreasonable to require that, after committing to achieving a certain goal, the intended states of affairs are preferable to its alternatives.

Lastly, by Cohen and Levesque's requirement 1, intentions pose problems for deliberation, meaning that intentions are only relevant if the agent believes they have not been achieved. Thus, this implies that

$$Int(\psi) \Rightarrow \neg B\psi.$$

The proposal of intentions as epistemically possible overwhelming desires does not encompass this requirement, thus we have to revise our proposal to consider only those desires which have not been achieved yet. Given the above discussion, we propose the following encoding of intention in our logic, generalizing it to a conditional form as has been proposed for both beliefs and goals:

$$Int(\psi|\varphi) \equiv I(\psi|\varphi) \wedge E(\psi \wedge \varphi) \wedge \neg B(\psi|\varphi) \tag{5.6}$$

In the following, however, we will extend the logic to represent the agent's capabilities to act in the world, and we will propose an extension of this encoding of intentions.

5.2.2 Intention and practicality

The relationship between intention and practicality is quite a different aspect than what we have been treating before. In our framework we do not have the machinery to represent ontic actions - i.e. actions that change the environment. To allow the representation of practicality, we must extend the language of $\mathscr{L}_{\leq_P,\leq_D}$ to incorporate ontic actions, or simply plans.

Definition 5.11 *We call $\mathscr{A} = \langle A, pre, pos\rangle$ an action library, or plan library, iff A is a set finite set of plans symbols, $pre : A \to \mathscr{L}_0$ is a function that maps each plan to a propositional formula representing its preconditions and $pos : A \to \mathscr{L}_0$ the function that maps each plan to a propositional formula representing its post-conditions. We further require that the post-conditions of any plan is a conjunction of propositional literals*[5]*. We say $\alpha \in \mathscr{A}$ for any plan symbol $\alpha \in A$.*

In this work we will use *plans* and *actions* interchangeably. Since we will not introduce complex action operators in our framework, such as composition, parallel execution etc., we believe this terminological decision does not affect the comprehension of our framework. We point out that, while formally we are introducing ontic actions in the framework, in the light of the philosophical inspirations for our construction, this actions are understood as the plans the agent may adopt to achieve a given situation.

We will include the plans in our language to create a logic of practical rationality we will use to define a more adequate notion of intention. For that, we will extend the language presented in Definition 5.1 to include formulas of the type $[\alpha]\varphi$ meaning "after the execution of α, φ holds".

Definition 5.12 *Let P be a set of propositional letters and $\mathscr{A}$ a plan library, we define the language $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$ by the following grammar (where $p \in P$ and $\alpha \in \mathscr{A}$):*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid A\varphi \mid [\leq_P]\varphi \mid [<_P]\varphi \mid [\leq_D]\varphi \mid [<_D]\varphi \mid [\alpha]\varphi$$

To model the effect of performing an ontic action $\alpha \in \mathscr{A}$ given an agent model M, we will define the notion of model update, as commonly used in the area of Dynamic Epistemic Logic.

[5] To simplify the formal machinery necessary to introduce ontic actions in our framework, we constrain the post-condition of plans to be conjunctive formulas. With that simplification, we can establish a simple model theory for actions, and by the framework of Van Ditmarsch and Kooi (2008), we can obtain a simpler axiomatization for this extended logic. From the point of view of Agent Programming, this is not a great constrain for the logic, since many such language impose similar restrictions for plans.

Definition 5.13 *Let $\mathscr{A} = \langle A, pre, pos\rangle$ be a plan library, $\alpha \in \mathscr{A}$ an action (or plan) and $M = \langle W, \leq_P, \leq_D, v\rangle$ an agent model. The product update of model M by action α is defined as the model $M \otimes [\mathscr{A}, \alpha] = \langle W', \leq'_P, \leq'_D, v'\rangle$ where*

$$\begin{array}{lcl} W' & = & \{w \in W \mid M, w \vDash pre(\alpha)\} \\ \leq'_P & = & \leq_P \cap\, W' \times W' \\ \leq'_D & = & \leq_D \cap\, W' \times W' \\ v'(p) & = & \begin{cases} W' & \text{if } pos(\alpha) \vDash p \\ \emptyset & \text{if } pos(\alpha) \vDash \neg p \\ v(p) \cap\, W' \times W' & \text{otherwise} \end{cases} \end{array}$$

Notice that, by the definition above, we are assuming the result of plans to be deterministic, i.e. always achieve the same result given the initial state. This definition is based on the definition of ontic actions of Van Ditmarsch and Kooi (2008).

From this we can define for any agent model M and a world w of M the conditions for satisfaction of a formula $[\alpha]\varphi$, meaning '*after the execution of a plan* α, it holds that φ'.

Definition 5.14 *Let $\mathscr{A}$ be a plan library and $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model.*

$$M, w \vDash [\alpha]\varphi \quad \textit{iff} \quad \textit{if } M, w \vDash pre(\alpha) \textit{ then } M \otimes [\mathscr{A}, \alpha], w \vDash \varphi$$

From the above definition, we provide the axiomatization of the augmented logic by including reduction axioms for plans to the axiom schemata presented in Figure 5.1. These reduction axioms are based on the axiomatization provided by Van Ditmarsch and Kooi (2008).

Proposition 5.15 *Let P be a set of propositional letters and $\mathscr{A}$ a plan library, the logic of $\mathscr{L}_{\leq_P, \leq_D}(P, \mathscr{A})$ is completely axiomatized by the axiom schemata presented in Figure 5.1 together with the reduction axioms presented in Figure 5.2, as well as the modus ponens and necessitation rules, for each $\alpha \in \mathscr{A}$.*

Regarding the relation between plans and beliefs, since the result of plans are deterministic, we have that after execution of the plan the agent believes in the effectivity of the plan, i.t. that the effects of the plan (its post-conditions) hold.

Figure 5.2 – Reduction Axioms for plans.

$$
\begin{array}{lcll}
[\alpha]p & \leftrightarrow & pre(\alpha) \rightarrow \top & \text{if } pos(\alpha) \vDash p \\
[\alpha]p & \leftrightarrow & pre(\alpha) \rightarrow \bot & \text{if } pos(\alpha) \vDash \neg p \\
[\alpha]p & \leftrightarrow & pre(\alpha) \rightarrow p & \text{if } pos(\alpha) \nvDash p \text{ and } pos(\alpha) \nvDash \neg p \\
[\alpha](\varphi \wedge \psi) & \leftrightarrow & [\alpha]\varphi \wedge [\alpha]\psi & \\
[\alpha]\neg\varphi & \leftrightarrow & pre(\alpha) \rightarrow \neg[\alpha]\varphi & \\
[\alpha][\leq_P]\varphi & \leftrightarrow & pre(\alpha) \rightarrow [\leq_P][\alpha]\varphi & \\
[\alpha][\leq_D]\varphi & \leftrightarrow & pre(\alpha) \rightarrow [\leq_D][\alpha]\varphi & \\
[\alpha][<_P]\varphi & \leftrightarrow & pre(\alpha) \rightarrow [<_P][\alpha]\varphi & \\
[\alpha][<_D]\varphi & \leftrightarrow & pre(\alpha) \rightarrow [<_D][\alpha]\varphi & \\
[\alpha]A\varphi & \leftrightarrow & pre(\alpha) \rightarrow A[\alpha]\varphi &
\end{array}
$$

Source: the author

Proposition 5.16 *Let $\mathscr{A}$ be a plan library and $\alpha \in \mathscr{A}$ a plan. In the logic of practical rationality, it is valid that*

$$B(pre(\alpha)) \quad \Rightarrow \quad [\alpha]B(pos(\alpha))$$

Proof: Immediate from the definition of B (formula 5.2) and Definition 5.14. □

With the addition of actions, we can represent the notion of ability, or that an agent **can** achieve a state of affairs s.t. φ holds. For that we will introduce the formula $\diamond\varphi$ meaning '*it is possible to achieve* φ.'

$$\diamond\varphi \equiv \bigvee_{\alpha \in \mathscr{A}} (pre(\alpha) \wedge [\alpha]\varphi) \tag{5.7}$$

Notice that the notion of achievability expressed above is quite simple, in the sense that a state of affairs is achievable iff it is a consequence of the execution of some plan. One could argue that the agent may compose her plans into a complex plan to achieve the desired property. This is, in fact, a valid point and there has been some work on using these kinds of ontic actions in DEL to model planning agents, as the work of Andersen, Bolander and Jensen (2014).

We must point out, however, that our work focuses on studying these phenomena from the perspective of Agent Programming. Since planning is a computationally costly task, most agent programming languages use pre-compiled plans or simple plan generation rules, which can be easily represented in our action libraries. As such, we claim that this encoding is, for the sake of studying agent programming languages, sufficiently expressive.

To model requirements 1, 5 and 6 of the Cohen and Levesque (1990)'s desiderata, if an agent intends ψ, than she must believe she can achieve ψ, i.e. there is a plan α, which the agent believes to be executable, thus $B(pre(\alpha))$ holds, and that after executing α, ψ holds, in other words $B([\alpha]\psi)$ holds. Well, this is exactly what we tried to encode with the formula $\Diamond\psi$, thus

to model these requirements, our notion of intention must satisfy:

$$Int(\psi) \Rightarrow B(\diamond\psi)$$

Now that we can express the ability of an agent to achieve a certain state of affairs by performing actions in its environment, we can amend our previous definition of intention to consider practicality.

$$Int(\psi|\varphi) \equiv I(\psi|\varphi) \wedge E(\varphi \wedge \psi) \wedge \neg B(\psi|\varphi) \wedge B(\diamond\psi|\varphi) \tag{5.8}$$

With the definition above, we revisit the desiderata for intentions of Cohen and Levesque and we conclude that:

1. Intentions pose problems for deliberations, i.e. an agent intends a state of affairs if, and only if, there is an action in her library to achieve that state of affairs;
2. Intentions pose a "window of acceptability" for other intentions, since $Int(\psi) \Rightarrow \neg Int(\neg\psi)$ is valid in the logic of rationality;
4. From the definition of intention $Int(\psi)$, it is clear that the agent must believe that the desired state of affairs is possible, since $E(\psi)$ and $B(\diamond\psi)$;
5. The agent does not believe she will not brig about her intentions, if $Int(\psi)$, it is clear that the agent must not believe that the desired state of affairs is unachievable, since $B(\diamond\psi)$ must hold;
6. Also, the agent does believe she will brig about her intentions, if the actual world comes to be one of the most plausible ones, since $Int(\varphi)$ implies $B(\diamond\psi)$, then there must be a plan $\alpha \in \mathscr{A}$, s.t. $B([\alpha]\psi)$ must hold;
7. The agent does not intend all the expected (believed) side effects of its intention, but, as it is the case for Cohen and Levesque's theory of intentions, all the non-believed logical implications of their intentions are also intended. In other words, $Int(\varphi) \wedge B(\varphi \rightarrow \psi) \not\Rightarrow Int(\psi)$ but $Int(\varphi) \wedge A(\varphi \rightarrow \psi) \wedge \neg B(\psi) \Rightarrow Int(\psi)$.

Notice that the main goal of our logic of practical rationality $\mathscr{L}_{\leq_P,\leq_D}$ is to represent the mental state of an agent in cognitive agent programming, not to characterize its execution. As such, our logic is not expressive enough to represent requirement 3, nor the aspects of requirements 5 and 6 related to action execution and failure, in Cohen and Levesque's desiderata for a theory of intention.

Intentions in our logic of practical rationality $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$, thus, represent that an agent **can adopt** an intention φ, in the sense that it satisfies all the desiderata for an intention, not that an agent **does** intend to φ. This will be an important difference in applying this logic to study agent programming semantics in Chapter 6.

5.3 Dynamics of mental attitudes in agent programming

In this section, we study the dynamic properties of the mental attitudes. From Chapter 4, we know that operations on the agent mental state can be presented by introducing dynamic operators in the language. From the connection provided by Theorem 5.7, and the reduction axioms and harmony results presented in Chapter 4, we can encode these operations on our logic $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$.

We will redefine the operations presented in Chapter 4 in our logic of practical rationality $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$. Since we now have two different preference relations, we will define the effect of each operation on each preference relation of the model. As such, we will achieve operations such as *plausibility update by radical upgrade* and *desirability update by public suggestion*, for instance.

5.3.1 Plausability and desirability update by public announcement

The first operation we study is public announcements. As already discussed in Chapter 4, public announcement is an epistemic operation that constrains the worlds the agent considers epistemically possible, by eliminating from the model all worlds in which the announced formula does not hold. It corresponds to the mental action of an agent acquiring more knowledge about the world. Below we define the transformation caused by public announcement operation over agent models as given in Definition 5.2. Definition 5.17 bellow is the counterpart of Definition 4.23 for these models.

Definition 5.17 *Let $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model and φ a formula of $\mathscr{L}_0$. We say the agent model $M_{!\varphi} = \langle W_{!\varphi}, \leq_{P!\varphi}, \leq_{D!\varphi}, v_{!\varphi}\rangle$ is the result of public announcement of φ in M, where:*

$$
\begin{aligned}
W_{!\varphi} &= \{w \in W \mid M, w \vDash \varphi\} \\
\leq_{P!\varphi} &= \leq_P \cap (W_{!\varphi}^2) \\
\leq_{D!\varphi} &= \leq_D \cap (W_{!\varphi}^2) \\
v_{!\varphi}(p) &= v(p) \cap W_{!\varphi}
\end{aligned}
$$

Analogous to Definition 4.24, we say $M, w \vDash [!\varphi]\psi$ iff $M, w \vDash \varphi$ implies $M_{!\varphi}, w \vDash \psi$.

Since the public announcement is a world removing operation, its occurrence affect both the plausibility and desirability relations of the agent. The axiomatization for this extended logic consists of the axiomatization presented in proposition 5.15, together with reduction axioms presented in Proposition 4.25 replicated for the modalities $[\leq_P]$ and $[\leq_D]$.

With this axiomatization and encodings proposed in the formulas 5.2, 5.4, 5.5 and 5.8, we can derive reduction axiom describing the changes in the mental attitudes of an agent after a public announcement. It is not so simple, however, to give an intuitive reading for these reduction axioms. This is because the mental attitudes *B*, *G*, *I* and *Int* are defined as complex formulas over the logic $\mathscr{L}_{\leq_P, \leq_D, \alpha}(P, \mathscr{A})$.

In our opinion, a much more fruitful way to present the changes in the mental state due to harmonic operations is using agent structures, given in Definition 5.6. In other words, we believe that using agent structures, we can obtain a more simple way of understanding changes performed in the agent's mental state after a certain mental action.

To provide such description of public announcements by means of agent structure, we define the notion of agent structure restriction, similar to the graph restriction presented in Definition 4.26. Since agent structures are composed by two priority graphs, the restriction of an agent structure $\mathscr{G}$ by a formula φ, denoted $\mathscr{G}^\varphi$ will be the restriction of each priority graph composing it.

Definition 5.18 *Let $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D \rangle$ be an agent structure and $\varphi \in \mathscr{L}_0$ a propositional formula. The restriction of $\mathscr{G}$ in regards to φ is the agent structure $\mathscr{G}^\varphi = \langle \mathscr{G}_P^\varphi, \mathscr{G}_D^\varphi \rangle$, where $\mathscr{G}_P^\varphi$ and $\mathscr{G}_D^\varphi$ are the graph restrictions of the priority graphs $\mathscr{G}_P$ and $\mathscr{G}_D$, respectively, as presented in Definition 4.26.*

With that, we can show how to describe the result of a public announcement operation over an agent model by means of its agent structure.

Proposition 5.19 *Let $M = \langle W, \leq_P, \leq_D, v \rangle$ be an agent model, $\varphi \in \mathscr{L}_0$ a propositional formula and $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D \rangle$ an agent structure, such that M is induced by $\mathscr{G}$. Then, $M_{!\varphi}$ is induced by $\mathscr{G}^\varphi$.*

Proof: Immediate from the Definition 5.17, and Theorems 5.7 and 4.27. □

5.3.2 Plausibility update by radical upgrade

The second operation we introduce for agent models is that of plausibility update. This operation corresponds to applying the radical upgrade presented in Chapter 4 to the plausibility structure of the model. This operation corresponds to an agent coming to (irrevocably) update her beliefs to accommodate a new belief.

Definition 5.20 *Let $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model and φ a formula of $\mathscr{L}_0$. We say hat $M_{\Uparrow_P \varphi} = \langle W, \leq'_P, \leq_D, v\rangle$ is the result of the plausibility update by radical upgrade of M by φ, where $\langle W, \leq'_P, v\rangle$ is the radical upgrade of $\langle W, \leq_P, v\rangle$ by φ as presented in Definition 4.28.*

We will use the notation $[\Uparrow_P \varphi]\psi$ to indicate that, after a plausibility update by radical upgrade of φ, ψ holds, as usual.

As before, the reduction axioms for plausibility update by radical upgrade consists of the reduction axioms presented in Proposition 4.30 for the modality $[\leq_P]$ together with the reduction axioms bellow. The following axioms are obtained observing that the plausibility update preserves the preference relation $\leq_D$, the set of possible worlds and the plan library of the language.

$$\begin{array}{lcl} [\Uparrow_P \xi][\alpha]\varphi & \leftrightarrow & [\alpha][\Uparrow_P \xi]\varphi, \text{ for each } \alpha \in \mathscr{A} \\ [\Uparrow_P \xi][\leq_D]\varphi & \leftrightarrow & [\leq_D][\Uparrow_P \xi]\varphi \\ [\Uparrow_P \xi][<_D]\varphi & \leftrightarrow & [<_D][\Uparrow_P \xi]\varphi \end{array}$$

Notice that, as we stated before, the reduction axioms for changes in the defined conditional modalities of beliefs, goals and intentions are not simple to give intuitive readings by means of changes in the agent's mental state. Seen as a transformation on an agent structure, we have that to update the plausibility order by means of a radical upgrade of φ amounts to perform a radical upgrade by φ on the agent's the belief base, i.e. on the priority graph describing her plausibility relation. Remind, from Proposition 4.32 that such operation amounts to prefix the priority graph by the singleton graph $\overline{\varphi} = \langle\{\varphi\}, \emptyset\rangle$. As such, we have the following result.

Proposition 5.21 *Let $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model, $\varphi \in \mathscr{L}_0$ a propositional formula and $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D\rangle$ an agent structure, such that M is induced by $\mathscr{G}$. Then, $M_{\Uparrow_P \varphi}$ is induced by $\mathscr{G} \Uparrow_P \varphi = \langle \overline{\varphi}; \mathscr{G}_P, \mathscr{G}_D\rangle$.*

5.3.3 Desirability update by radical upgrade

Similarly to the previous action, we define the case of an update on the agent's preference relation $\leq_D$. This action means that the agent adopts the goal to φ - or, in another words, revise her goal base to include φ.

Definition 5.22 *Let $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model and φ a formula of $\mathscr{L}_0$. We say that $M_{\Uparrow_D \varphi} = \langle W, \leq_P, \leq_{D\Uparrow\varphi}, v\rangle$ is the result of the desirability update by radical upgrade of M by φ, where $\langle W, \leq_{D\Uparrow\varphi}, v\rangle$ is the radical upgrade of $\langle W, \leq_D, v\rangle$ by φ as presented in Definition 4.28.*

We will use the notation $[\Uparrow_D \varphi]\psi$ to indicate that, after a desirability update by radical upgrade of φ, ψ holds.

As before, the reduction axioms for desirability update by radical upgrade consists of the reduction axioms presented in Proposition 4.30 for the modality $[\leq_D]$ together with the reduction axioms bellow. As for plausibility update, the following axioms were obtained observing that the desirability update by radical upgrade preserves the plausibility relation.

$$[\Uparrow_D \xi][\alpha]\varphi \leftrightarrow [\alpha][\Uparrow_D \xi]\varphi, \text{ for each } \alpha \in \mathscr{A}$$
$$[\Uparrow_D \xi][\leq_P]\varphi \leftrightarrow [\leq_P][\Uparrow_D \xi]\varphi$$
$$[\Uparrow_D \xi][<_P]\varphi \leftrightarrow [<_P][\Uparrow_D \xi]\varphi$$

Seen as a transformation on an agent structure, we have that to update the desirability relation by means of a radical upgrade of φ amounts to prefixing the agent's goal base, i.e. the priority structure representing her desirability relation $\leq_D$, by $\overline{\varphi}$. As such, we obtain the following representation result.

Proposition 5.23 *Let $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model, $\varphi \in \mathscr{L}_0$ a propositional formula and $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D\rangle$ an agent structure, such that M is induced by $\mathscr{G}$. Then, $M_{\Uparrow_D\varphi}$ is induced by $\mathscr{G} \Uparrow_D \varphi = \langle \mathscr{G}_P, \overline{\varphi}; \mathscr{G}_D\rangle$.*

5.3.4 Desirability update by public suggestion

As before, we will redefine the operation of public suggestion on agent models. Here we will present the desirability update by public suggestion. The plausibility update by public suggestion can be defined similarly exchanging $\leq_D$ and $<_D$ for $\leq_P$ and $<_P$ in the definitions below.

Definition 5.24 *Let $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model and φ a formula of $\mathscr{L}_0$. We say that $M_{\#_P\varphi} = \langle W, \leq_P, \leq_{D\#\varphi}, v\rangle$ is the result of the desirability update of M by public suggestion of φ, where $\langle W, \leq_{D\#\varphi}, v\rangle$ is the result of the suggestion of φ in $\langle W, \leq_D, v\rangle$ as presented in Definition 4.33*

We will use the notation $[\#_D\varphi]\psi$ to indicate that, after a desirability update by radical upgrade of φ, ψ holds.

As before, the reduction axioms for desirability update by public suggestion consist of the reduction axioms presented in Proposition 4.35 for the modality $[\leq_D]$ together with the reduction axioms bellow. The axioms bellow were obtained noticing that the desirability update by public suggestion preserves the plausibility relation, the set of possible worlds and the plan library.

$$[\#_D\xi][\alpha]\varphi \leftrightarrow [\alpha][\#_D\xi]\varphi, \text{ for each } \alpha \in \mathscr{A}$$
$$[\#_D\xi][\leq_P]\varphi \leftrightarrow [\leq_P][\#_D\xi]\varphi$$
$$[\#_D\xi][<_P]\varphi \leftrightarrow [<_P][\#_D\xi]\varphi$$

Seen as a transformation on an agent structure, we have that to update the desirability relation by means of a public suggestion of φ amounts to the parallel composition of the agent's goal base, i.e. the priority graph $\mathscr{G}_D$ describing the agents desirability relation, and the singleton graph $\overline{\varphi}$, as presented in Definition 4.36.

Proposition 5.25 *Let $M = \langle W, \leq_D, \leq_P, v\rangle$ be an agent model, $\varphi \in \mathscr{L}_0$ a propositional formula and $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D\rangle$ an agent structure, such that M is induced by $\mathscr{G}$. Then, $M_{\#_D\varphi}$ is induced by $\mathscr{G}\#_D\varphi = \langle \mathscr{G}_P, \overline{\varphi} \parallel \mathscr{G}_D\rangle$.*

5.3.5 Plausibility update by contraction

We now define the contraction of a belief from the plausibility ordering. This operation corresponds to applying the lexicographic contraction presented in Chapter 4 to the plausibility relation of the model. This operation corresponds to an agent contracting a belief.

Definition 5.26 *Let $M = \langle W, \leq_P, \leq_D, v\rangle$ be an agent model and φ a formula of $\mathscr{L}_0$. We say that $M_{\Downarrow_P\varphi} = \langle W, \leq_{P\Downarrow\varphi}, \leq_D, v\rangle$ is the result of the the plausibility update of M by contracting φ, where $\langle W, \leq_{P\Downarrow\varphi}, v\rangle$ is the lexicographic contraction of φ from $\langle W, \leq_P, v\rangle$ as presented in Definition 4.48.*

We will use the notation $[\Downarrow_P \varphi]\psi$ to indicate that, after a plausibility update by contraction of φ, ψ holds, as usual.

As before, the reduction axioms for plausibility update by contraction consists of the reduction axioms presented in Proposition 4.50 for the modality $[\leq_P]$ together with the reduction axioms bellow, obtained by preservation of the relation $\leq_D$ as before.

$$[\Downarrow_P \xi][\alpha]\varphi \leftrightarrow [\alpha][\Downarrow_P \xi]\varphi, \text{ for each } \alpha \in \mathscr{A}$$
$$[\Downarrow_P \xi][\leq_D]\varphi \leftrightarrow [\leq_D][\Downarrow_P \xi]\varphi$$
$$[\Downarrow_P \xi][<_D]\varphi \leftrightarrow [<_D][\Downarrow_P \xi]\varphi$$

The result of contracting the formula φ from the agents beliefs corresponds to the removal of φ from the agent's belief base. It is modelled by the syntactic transformation of the priority graph representing the agent's plausibility relation, by means of the construction presented in Definition 4.59.

Proposition 5.27 *Let $M = \langle W, \leq_D, \leq_P, v\rangle$ be an agent model, $\varphi \in \mathscr{L}_0$ a propositional formula and $\mathscr{G} = \langle \mathscr{G}_P, \mathscr{G}_D\rangle$ an agent structure, such that M is induced by $\mathscr{G}$. Then, $M_{\Downarrow_P \varphi}$ is induced by $\mathscr{G} \Downarrow_P \varphi = \langle \mathscr{G}_P \Downarrow \varphi, \mathscr{G}_D\rangle$.*

A similar definition can be provided for desirability update by a contraction changing $\leq_P$ for $\leq_D$ and $<_P$ for $<_P$ in the definitions above.

5.4 Summary of the chapter

In this chapter, we use the Dynamic Preference Logic presented before to encode mental attitudes. We begin describing a logic for rationality or an agent specification logic in Section 5.1 which is used to encode the notion of belief and goal (equivalent to that of desire, for our purposes). In Section 5.2, we analyse an encoding of the notion of intention in the logic $\mathscr{L}_{\leq_P, \leq_D}$ of plausibility and desirability and, giving the requirements for a formalization of intention discussed in Chapters 2 and 3, we propose an extension of $\mathscr{L}_{\leq_P, \leq_D}$ to represent agent's plans. With this extended logic, we encode the notion of intention. Finally, in Section 5.3, this logic was be dynamified, in a way similar to what has been done in the previous chapter.

Notice that our definition satisfy Bratman's strong consistency requirements. Also, by our definition, an agent reconsiders an intention that φ if, and only if, (i) some alternative to φ become as equally attractive; (ii) the agent does not believe φ to be possible or (iii) the

agent does not believe φ is achievable. In other words, our intentions satisfy Mintoff (2004)'s principles for reconsideration.

Part III

The Application

6 AN ABSTRACT PROGRAMMING LANGUAGE FOR AGENTS

The aim of this chapter is to show that the logic of practical rationality $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$ presented in Chapter 5 contains the fundamental elements present in most BDI agent programming language. Also, we wish to show that, by means of agent structures presented in Definition 5.6, we can provide a declarative interpretation of the mental state of agent programs for many different programming languages.

We will show how to translate an agent program state, as defined in the semantics of the programming language, into an equivalent agent structure, as presented in Definition 5.6. Since the agent program states we will define for the abstract agent programming language proposed in this chapter are commonly used in the semantics of several BDI-inspired programming languages, we believe the discussions and results presented in this chapter are of general appeal to the area of Agent Programming.

Further yet, we propose that if we can specify the formal semantics of the programming language by means of some common operations on the agent program, we can connect the changes in the agent's program state to the dynamic operations studied in Chapter 4 and 5. On other words, since the operational semantics of the programming language we propose is given by means of mental operations such as addition and removal of beliefs, goals and intention, if we can guarantee these operations are harmonic, this semantics can be subsumed by the declarative interpretation of the changes performed in the mental state of the agent.

In the following, we introduce in Section 6.1 the abstract agent programming language AAP. In Section 6.2, we define an abstract operational semantics which will be realized into concrete interpreter by means of auxiliary selection functions in Section 6.3. In Section 6.4, we provide the connection between the language AAP and the logic presented in Chapter 5, showing how a program state in this programming language can be interpreted as an agent structure defined in Chapter 5 and how the programming language semantics can be connected to the dynamic operations studied in that chapter. Finally, we analyse the semantics of our language in contrast to the philosophical requirements studied in Chapter 2 in Section 6.5.

6.1 The AAP language

In this section, we propose an abstract agent programming language AAP. In AAP, an agent program is composed of a structure $ag = \langle K, B, G, I \rangle$ describing the agent's mental state, similar to many BDI-inspired agent programming languages in the literature, defined over a

plan library describing the plans available for the agent to achieve her goals.

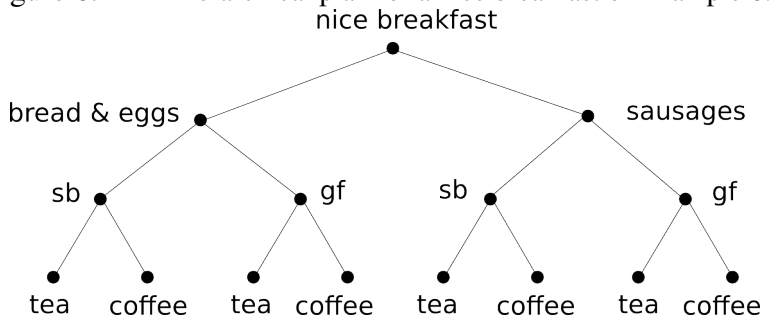
In the definition of our language, we will follow Bratman (1999)'s intentions-as-plans paradigm, in which an agent's intentions are composed of the plans describing how to achieve the desired goals. As such, the presentation of the elements of our language begins with the definition of what is a plan in this context.

As Bratman (1999) points out, plans may be partially specified, in the sense that they need to be further refined until they become concrete, i.e. detailed enough in the current state of affairs to be executed. Only concrete plans may be executed, since by definition partial plans need to be further specified before executed.

We will model this notion of a partial plan being refined by the notion of a hierarchical plan. Informally, a hierarchical plan is a finite AND-OR tree such that each node of the tree possess a pre-condition and a post-condition. Each node of the hierarchical plan describes a certain level of refinement from the original abstract plans- represented by the root of the tree. As such, the descendants of any node in the tree can be understood as alternative refinements of the plans represented by the node. Let's examine Example 6.1.

Example 6.1 *An agent intends to have a nice breakfast before leaving for work. A nice breakfast for her, is composed of some fresh fruit , a warm beverage, and either bread and eggs or some breakfast meat. In her kitchen, the agent has fruits such as strawberries (sb) and grapefruit (gf), tea and coffee, bread, eggs and sausages. A representation of a hierarchical plan for the agent to have a nice breakfast is depicted in Figure 6.1. There, the tree can be read as: to have a nice breakfast, the agent may have either bread and eggs or sausages. To have a nice breakfast with bread and eggs, the agent can have either strawberries or grapefruit, etc.*

Figure 6.1 – A hierarchical plan for a nice breakfast of Example 6.1



Source: the author

This notion of plan refinement is commonly implemented in agent programming languages, such as AgentSpeak (RAO, 1996), by iterative selection of plans to achieve subgoals.

As such, an hierarchical plan, as presented in Example 6.1, constitutes an abstract representation of all the forms a plan can be achieved. Similar representations of hierarchical plans appear in the literature, for example, in the work of plan revision by Van der Hoek, Jamroga and Wooldridge (2007). We now formalize this notion of hierarchical plans.

Definition 6.2 (Hierarchical Plan) *Let P be a set of propositional symbols. A hierarchical plan on P, or simply hierarchical plan, is a tree*[1] *$\pi = \langle N, \prec, pre, pos\rangle$, where N is a finite set of nodes, $\prec \subseteq N \times N$ a pre-order over N, $pre : N \to \mathscr{L}_0(P)$ and $pos : N \to \mathscr{L}_0(P)$ are the pre and post-condition functions. We further require that:*

- *For any $i \in N$, $pre(i) \nvdash \bot$ and $pos(i) = l_1 \wedge \ldots \wedge l_n \nvdash \bot$, where l_i is a propositional literal;*
- *If $i, j \in N$ s.t. $i \prec j$ then $pre(i) \vDash pre(j)$ and $pos(j) \vDash pos(i)$;*

The first requirement in the Definition 6.2 is that the hierarchical plan is a tree, i.e. the structure $\langle N, \prec\rangle$ has a minimal element - the root of the tree - and each node has at most one parent.

Second, we require that plans are internally consistent, i.e. both its pre-conditions and post-conditions are consistent, and that the post-condition of plans are described by a conjunction of propositional literals. This requirement guarantees that plans in the language AAP can be modelled, in the logical level, by ontic actions as defined in Chapter 5. Also, requiring the post-conditions to be specified by a conjunction of literals, we can reduce complexity in reasoning about the consequences of plans. This is not a serious limitation on the expressibility of our language since this is a common requirement in agent programming languages.

The final requirement guarantees that the relation $\prec$ has the intended meaning of refinement, or further specification of a plan. Given two nodes i, j, if we have that $i \prec j$, the intuition we presented requires that j refines i. As such, any situation in which the plan i is applicable, plan j must also be. Further yet, if j is a refinement of i, then accomplishing j must also accomplish i. Through the refinement of plans we will model, in Section 6.2, the reasoning process of an agent making decisions to achieve her goals.

Any node in a hierarchical plan describes the path from the root to that node, corresponding to a set of choices made by the agent about how to pursue a given goal. Given a hierarchical plan $\pi = \langle N, \prec, pre, pos\rangle$ and a node n, we can thus describe the path from the root to n. As such, we will represent a path in π reaching n simply by the pair $\langle \pi, n\rangle$.

[1] A tree is a partially ordered set $\langle T, \prec\rangle$, such that there is a minimum element in T and for any $t \in T$ the set $\downarrow(t) = \{s \in T \mid s \prec t\}$ is well ordered.

Let's define some terminology for us to simplify the description of the behaviour of the agents using hierarchical plans. This terminology describe structural notions that will be important for us, that of *root*, *successor* and *leaf* of a hierarchical plan.

Definition 6.3 *Let $\pi = \langle N, \prec, pre, pos\rangle$ be a hierarchical plan and $r \in N$ its $\prec$-minimal element. We call r the root of π and denote such element by $root(\pi)$.*

We will also need the notion of successor of a node. Informally, a successor of a node n represents a direct refinement of the plan represented by the node n. In Example 6.1, a successor of the element *bread & eggs* is the element *gp*, meaning that a *nice breakfast with bread and eggs and grapefruit* is a refinement of the plan to have a *nice breakfast with bread and eggs*.

Definition 6.4 *Let $\pi = \langle N, \prec, pre, pos\rangle$ be a hierarchical plan and $i, j \in N$ two nodes. We say j is a refinement of plan of i, denoted by $j \in suc(i)$, iff $i \prec j$ and there is no k, s.t. $i \prec k \prec j$.*

The notion of leaf will also be of use for us. The leafs of a hierarchical plan represent the concrete plans of the agent, i.e. those plans that have been completely specified. In Example 6.1, the lowermost and leftmost node represents the concrete plan to have *a nice breakfast by having bread and eggs, strawberries and tea*.

Definition 6.5 *Let $\pi = \langle N, \prec, pre, pos\rangle$ be a hierarchical plan and $i \in N$ a node. We say that i is a concrete plan in π, $i \in leaf(\pi)$, if i has no successor, i.e. $suc(i) = \emptyset$.*

We will call any (finite) set of hierarchical plans Π an AAP plan library[2].

Definition 6.6 *Let Π be a finite set of hierarchical plans, we call Π an AAP plan library. Also, we denote by $\mathscr{H}_\Pi$ the set of paths in the hierarchical plans of the plan library Π, representing all the (possibly partially specified) plans available to the agent.*

Let's consider our Example 6.1. Let's encode one path of Figure 6.1, giving its pre-conditions and post-conditions

Example 6.7 *To have a nice breakfast she must have some fresh fruit, a warm beverage, and either bread and eggs or sausages. We can provide an hierarchical plan π, s.t. each node n of π can be described by a plan of the form*

$$\rho : \rho' = \varphi \leftarrow \psi,$$

[2]This is not the same as the notion of plan library presented in Chapter 5. The connection between these two notions will be explored in Section 6.4

where ρ is the name of the (partial) plan, ρ' the antecessor of ρ in π, i.e. ρ' is a plan in π and $\rho \in suc(\rho')$. The path from the root to the leaf encoding the plan to have a nice breakfast constituted of bread, eggs, fresh strawberries and hot coffee can be encoded as the plans:

$$\begin{aligned}
\rho_0 : \top &= morning \wedge hungry \leftarrow nicebf \\
\rho_1 : \rho_1 &= morning \wedge hungry \leftarrow bread \ \& \ eggs, \ fresh \ fruit, \ warm \ beverage \\
\rho_2 : \rho_2 &= morning \wedge hungry \leftarrow bread \ \& \ eggs, \ strawberries, \ warm \ beverage \\
\rho_3 : \rho_3 &= morning \wedge hungry \leftarrow bread \ \& \ eggs, \ strawberries, \ coffee
\end{aligned}$$

Van der Hoek, Jamroga and Wooldridge (2007) shows also how to encode such plans as decision rules such as:

$$\begin{aligned}
nice\,breakfast &\leftarrow bread \ \& \ eggs, \ fresh \ fruit, \ warm \ beverage \\
fresh \ fruit &\leftarrow strawberries \\
warm \ beverage &\leftarrow coffee
\end{aligned}$$

Finally, we can define an abstract agent in our language.

Definition 6.8 *Let P be a set of propositional variables and Π a plan library. We call an agent program (an agent state or even an agent program state) over Π, a tuple $ag = \langle K, B, G, I \rangle$ where:*

- *$K \subset \mathscr{L}_O(P)$, is a consistent finite set of propositional formulas called the knowledge base;*
- *$B \subset \mathscr{L}_O(P) \times \mathbb{N}^*$ is a finite set of pairs $\langle \varphi, i \rangle$, called a stratified belief base, where φ is a propositional formula and i is a natural number, called the plausibility or rank of φ in B.*
- *$G \subset \mathscr{L}_O(P) \times \mathbb{N}^*$ is a finite set of pairs $\langle \varphi, i \rangle$, called a stratified goal base, where φ is a propositional formula and i is a natural number, called the desirability or rank of φ in G.*
- *$I \subset \mathscr{H}_\Pi$ is a finite set of paths $\langle \pi, n \rangle$ where $\pi \in \Pi$, called the (procedural) intention base.*

When the plan library Π is clear, we will often call the tuple $ag = \langle K, B, G, I \rangle$ an agent program (agent state or agent program state).

Notice above that the definition of belief and goal bases as pairs $\langle \varphi, i \rangle$ encode the notion that an agent believes (desires) that φ with a plausibility (desirability) of i. As in Chapter 5, the lower the plausibiliry (desirability) rank, the higher the plausibility (desirability) of a formula.

This description of belief and goal bases is equivalent to the representation of the stratified bases commonly used in belief base change (ROTT, 2009). A stratified belief base in

the area of belief base change is a totally ordered set of sets of propositional formulas $\Gamma = \langle h_1, \ldots, h_n \rangle$, where $h_i \subseteq \mathscr{L}_0$. The index i of each propositional set h_i denotes the plausibility of the formulas in the set h_i. As such, if $i < j$ we have that the formulas in h_i are more plausible than the formulas in h_j. It is clear that our representation of B and G can be converted into stratified bases by taking $h_i = \{\varphi \mid \langle \varphi, i \rangle \in B\rangle\}$. These bases correspond to a generalization of the flat bases (or sets of sentences) commonly used in several programming languages, such as AgentSpeak (RAO, 1996), GOAL (HINDRIKS, 2008) and 3APL (DASTANI; VAN RIEMSDIJK; MEYER, 2005).

More yet, since a path in a hierarchical plan represents a (possibly partially specified) plan, requiring the intention base I to be composed of paths, i.e. $I \subset \mathscr{H}_\Pi$, indicates our adherence to Bratman (1999)'s intention as plans paradigm. As such, we call any plan $\langle \pi, n \rangle \in I$ a procedural intention.

Example 6.9 *We can provide an AAP program for the breakfast agent of Example 6.1 as the agent* $ag = \langle K, B, D, I \rangle$ *over the plan library containing the hierarchical plan* π *discussed in Example 6.7, where* $B = \{\langle Morning, 1\rangle, \langle Hungry, 2\rangle\}$, $D = \{\langle Nicebf, 1\rangle\}$, $I = \{\langle \pi, \rho_0 \rangle\}$ *and* ρ_0 *is as defined in Example 6.7, and*

$$K = \left\{ \begin{array}{l} bread \ \& \ eggs \wedge fresh \ fruit \wedge \ warm \ beverage \rightarrow nicebf, \\ strawberries \rightarrow \ fresh \ fruit, \\ coffee \rightarrow \ warm \ beverage \end{array} \right\}.$$

It will be useful to refer to the preconditions and post-conditions of the plans in the intention base I. As such, we introduce some terminology.

Definition 6.10 *Let* $ag = \langle K, B, G, I \rangle$ *be an AAP agent program. We call the set of preconditions of the intention base I the set*

$$pre(I) = \{pre(n) \mid \langle \pi, n \rangle \in I\}$$

Additionally, we call the set of post-condtitions of the intention base I the set

$$pos(I) = \{pos(n) \mid \langle \pi, n \rangle \in I\}$$

6.1.1 Mental attitudes in AAP

Once established the main notion of an AAP agent program, we must now discuss how to interpret such a program in the context of Agent Programming. In other words, we must provide the computational interpretations of mental attitudes distilled in these programs.

An agent program over a plan library Π completely describes the mental state of an agent in the language AAP. As such, we need to define what it means for an agent represented by such agent program to know, belief, desire or intend a given state of affairs. Let's start with the notion of *knowledge*.

The knowledge base K in an agent program describes all the knowledge the agent possess. Knowledge for us is an undefeasible true belief, in the sense that if an agent knows that φ, then φ must be true and not falsifiable in the agent's mental state. Since knowledge is true belief and logical omniscience is not a concern for our computational agents, we require that an AAP agent knows everything that can be deduced from her knowledge base K.

Definition 6.11 *Let $ag = \langle K, B, G, I \rangle$ be an AAP agent program and $\varphi \in \mathscr{L}_0$. We say agent 'ag knows φ', denoted by $ag \models K\varphi$, iff $K \models \varphi$. We call $Know(ag) = \{\varphi \in \mathscr{L}_0 \mid K \models \varphi\}$ the knowledge set of agent ag.*

As we have pointed out, belief and goal bases are similar to stratified belief bases commonly used in Belief Base Revision. In this area, a way to the extrapolate of belief base to belief set has already been proposed. For example, given a stratified belief base $\Gamma = \langle h_1, \ldots, h_n \rangle$, Rott (2009) defines the belief set of an agent as the Tarskian consequence (given by Cn below) of the maximal consistent prefix of Γ, i.e.

$$Bel(\Gamma) = Cn(\bigcup_{1}^{i} h_i) \qquad \text{where i is s.t. } \bigcup_{1}^{i} h_i \nvdash \bot \text{ and } \bigcup_{1}^{i+1} h_i \models \bot$$

Notice, however, that limiting oneself to the maximal consistent *prefix*, one looses quite a bit of information. Examine the following case:

$$B = \{\langle p, 1 \rangle, \langle q, 2 \rangle, \langle \neg p, 3 \rangle, \langle r, 4 \rangle\}$$

the maximal consistent prefix of the belief base B above is the set $\{p, q\}$

From the maximal consistent prefix, however, we cannot derive the information r, despite the fact that it is explicitly stated in the belief base B and r is not inconsistent with any information more plausible than it. If we are to take plausibility (or rank) information seriously,

we can overcome the fact that an element $\neg p$ is stored in the belief base since it is inconsistent with the information p which is more plausible, without discarding *all* information less plausible than $\neg p$.

Hence, we define the consequence of a stratified belief base as the consequence of a maximal subset of the stratified base, respecting the plausibility ordering.

Definition 6.12 *Let $\Gamma \subset \mathscr{L}_0 \times \mathbb{N}$ be a finite set of pairs $\langle \varphi, i \rangle$ and let $\Gamma_i = \{\varphi \mid \langle \varphi, i \rangle \in B\}$. We define the maximal consistent subset of Γ, the set $\Gamma^{Max} \subset \mathscr{L}_0$, s.t.*

- $\Gamma^{Max} \subseteq \bigcup \Gamma_i$ *and if* $\langle \varphi, i \rangle \in \Gamma$ *and* $\varphi \in \Gamma^{Max}$ *then* $\Gamma_i \subseteq \Gamma^{Max}$;
- $\forall \Gamma' \subseteq \Gamma : (\exists \Gamma_i \subseteq \Gamma' \wedge \Gamma_i \not\subseteq \Gamma^{Max} \Rightarrow \Gamma' \vDash \bot$ *or* $\exists \Gamma_j \subseteq \Gamma^{Max} \wedge \Gamma_j \not\subseteq \Gamma'$ *and* $j < i)$

The maximal consistent subset of a stratified base consists of discarding only the *strata* which are incompatible with the information more plausible than it. The first requirement in Definition 6.12 guarantees that the entire *stata*, i.e. all formulas having a same rank, are incorporated into the maximal consistent subset. The reason for such condition is that, otherwise, we would have multiple possibilities of maximal subsets.

Revisiting the belief base B presented above, we obtain:

$$B^{Max} = \{p, q, r\}.$$

Notice that the knowledge held by an agent constrains the beliefs she can entertain as plausible, simply because her knowledge represents all the *truthful* and *indisputable* information she holds about the world. As such, when representing what the agent *believes* about the world, we must also consider what she *knows* about the world. To represent how an agent's knowledge base K contributes to her beliefs, we define the expanded belief base B_K of an agent.

Definition 6.13 *Let $ag = \langle K, B, G, I \rangle$ be an agent program. We call the extended belief base of ag the stratified base:*

$$B_K = \{\langle \varphi, 0 \rangle \mid \varphi \in K\} \cup B$$

With this notion, we define the belief set of an agent.

Definition 6.14 *Let $ag = \langle K, B, G, I \rangle$ be an agent program and $\varphi \in \mathscr{L}_0$. We say 'agent ag believes in φ', denoted by $ag \vDash B\varphi$, iff $B_K^{Max} \vDash \varphi$. We denote by $Bel(ag) = Cn(B_K^{Max})$ the belief set of agent ag.*

Example 6.15 *The belief set of the agent ag, with belief base $B = \{\langle Morning, 1 \rangle, \langle Hungry, 2 \rangle\}$, as described in Example 6.9, is the set $Bel = Cn(K \cup \{Morning, Hungry\})$.*

From Chapters 3 and 5 we know that Beliefs and Goals behave quite differently. Particularly, as we have claimed in Chapter 3, we base our encoding of goals (or desires) in that of Van Riemsdijk, Dastani and Meyer (2009). Remember that for Van Riemsdijk, Dastani and Meyer (2009), an agent has a goal to φ if there is a consistent subset of her goal base that implies φ.

It is not clear how we can generalize their encoding to stratified bases. What is clear in their encoding, however, is that goals are existential in nature, while beliefs are universal. Given that an agent holds a belief if it is a valid consequence of the maximal consistent subset of her extended belief base, we model the dual nature of goals and beliefs by requiring that an AAP agent has φ as a goal, if φ is consistent with the maximal consistent subset of her goal base.

As before, notice that the knowledge held by an agent constrains what the agent considers as possible. As such, when representing what the agent *desires* about the world, we must also consider what she *knows* about the world - agents in AAP are, thus, *realists* in the sense they cannot desire for what they know to be impossible. We do not believe this to be a burden, since we are modelling rational agents. To represent how an agent's knowledge base K influences her goals, we define the expanded goal base G_K of an agent.

Definition 6.16 *Let $ag = \langle K, B, G, I \rangle$ be an agent program. We call the extended goal base of ag the stratified base:*

$$G_K = \{\langle \varphi, 0 \rangle \mid \varphi \in K\} \cup G$$

We can, thus, define the notion of *goal* in AAP.

Definition 6.17 *Let $ag = \langle K, B, G, I \rangle$ be an agent program and $\varphi \in \mathscr{L}_0$. We say that 'the agent ag desires φ', denoted $ag \vDash G\varphi$, iff $G_K^{Max} \cup \{\varphi\} \nvdash \bot$. We denote by $Goal(ag) = \{\varphi \in \mathscr{L}_0 \mid ag \vDash G\varphi\}$ the goal set of the agent ag.*

Returning to our running example, we can analyse the desires of our breakfast agent

Example 6.18 *The goal set of the agent ag, with belief base $G = \{\langle Nicebf, 1 \rangle\}$, as described in Example 6.9, is the set $Goal = \{\varphi \mid K \cup \{Nicebf\} \nvdash \neg\varphi\}$.*

As briefly discussed before in Definition 6.8, since we adopt the intention as plans paradigm, intentions are composed of the plans the agent adopted to achieve them - described by the intention base I. Besides that procedural description, an intention also has a declarative aspect concerning its relation with the goal the agent intends to achieve.

In our hierarchical plans, this connection between a plan and a goal is described by the plan's post-condition. An intended goal, thus, is a goal ($\varphi \in Goal(ag)$) that has been adopted

to be achieved, i.e. a goal for which the agent has adopted a plan ($\langle \pi, n \rangle \in I$) to achieve it ($pos(n) \models \varphi$).

As such, we provide the following codification of *declarative* intentions.

Definition 6.19 *Let $ag = \langle K, B, G, I \rangle$ be an agent program and $\varphi \in \mathscr{L}_0$. We say 'the agent ag intends φ', denoted by $ag \models I\varphi$, iff $ag \models G\varphi$ and $\exists \langle \pi, n \rangle \in I$, s.t. $pos(n) \models \varphi$. We denote by $Int(ag) = \{\varphi \in \mathscr{L}_0 \mid ag \models I\varphi\}$ the (declarative) intention set of ag*

Example 6.20 *The intention set of the agent ag, with procedural intention base $I = \{\langle \pi, \rho_0 \rangle\}$, $pre(\rho_0) = Morning \wedge Hungry$ and $pos(\rho_0) = Nicebf$, as described in Example 6.9, is the set $Int = Cn(\{Nicebf\})$.*

Now that we have an interpretation of the declarative notions of belief, desire and intention in abstract agent programs, we can define the consistency requirements for the agent's mental state, based on Bratman (1999)'s strong consistency requirements.

Definition 6.21 *Let $ag = \langle K, B, G, I \rangle$ be an agent program. We say ag is coherent iff all of the conditions below hold.*

1. *for any path $\langle \pi, n \rangle \in I$ representing a plan, there is a goal $\varphi \in Goal(ag)$, such that $\langle \pi, n \rangle$ is a means to φ, i.e. $pos(n) \models \varphi$, i.e. intended states are desired states;*
2. *the plans of the agent are pursuable: $\forall \varphi \in pre(I)$, $Bel(ag) \models \varphi$;*
3. *the plans of the agent are consistent: $pos(I) \not\models \bot$, i.e. intentions are jointly-consistent.*
4. *the plans of the agent are relevant: $\forall \varphi \in pos(I)$, $Bel(ag) \not\models \varphi$;*

Now that we have defined the basic structure of a coherent abstract agent program, we will present the semantics of the proposed language.

6.2 Semantics of AAP

In this section, we propose an operational semantics for the language AAP. This semantics is formally defined as a transition relation between agent states previously defined in Definition 6.8.

The execution of an agent program will be defined by presenting a set of rules specifying how an agent state changes as a result of the execution of a certain action. The actions we will model are mental state changes corresponding to ontic or internal actions performed by the

agents. These kinds of actions are a common occurrence in agent programming languages, such as the perception, detection of events, plan selection and plan execution.

We propose, in this section, an abstract semantics for the language, in the sense that the transition relation defining our semantics is non-deterministic. In Section 6.3, we show how we can provide deterministic transition relation by means of what we call selection function. The reason behind this choice of presentation is to clearly separate the semantics of the language from any implementation choices. Also, by this separation, our semantics can be specialized to represent the reasoning cycle of different agent programming languages, without the need to revise the semantics.

Let's first introduce formally the actions we will use to represent the agents' deliberations.

Definition 6.22 *Let $ag = \langle K, B, G, I \rangle$ be an agent program. We define the allowed actions of agent ag as the set Ac_{ag} containing her intentions, as well as mental actions for addition of a new knowledge φ ($k\varphi$) or the addition ($b^+\varphi$) or removal ($b^-\varphi$) of beliefs, goals ($g^+\varphi$ or $g^-\varphi$) and intentions ($i^+\varphi$ or $i^-\varphi$), for any propositional formula φ:*

$$Ac_{ag} = I \cup \{k\varphi, b^+\varphi, b^-\varphi, g^+\varphi, g^-\varphi, i^+\varphi, i^-\varphi \mid \varphi \in \mathscr{L}_0\}.$$

The allowed actions describe in any given point what an agent can do. The allowed action are composed of the agent's plans or mental actions describing changes in the agents mental states.

Now we can formally specify the transition relation $\longrightarrow$ that will define the formal semantics of our abstract language AAP. Let S be the set of all agent program states, we will define the semantic transition function $\longrightarrow$ for AAP as the smallest relation $\longrightarrow \subseteq (S \times \bigcup_{ag \in S} Ac_{ag}) \times S$ satisfying the rules presented below. From that we conclude that an agent in AAP at any point in time can only do what she intends to do or change her mind about something.

Let $ag, ag' \in S$ and $a \in Ac_{ag}$, in the description of the rules below, we denote that $\langle ag, a \rangle \longrightarrow ag'$ by saying the if $\mathscr{M} = a$ then $ag \longrightarrow ag'$. The notation $\mathscr{M} = a$ is used to simulate a step function, which will be discussed later in Section 6.3. In this presentation, however, this notation is not but a stylistic choice made firstly to simplify the formal description of the rules and second, to make them compatible with the specification of any concrete interpreter, as discussed in Section 6.3.

6.2.1 Executing a plan

Notice that we are making some simplifying assumptions here. First, we suppose the execution of a plan is done in its entirety and all at once, i.e. any believed consequence of the plan is only achieved after the complete execution of the plan. This means that we are not considering actions as distributed over time, since this would require the information of time necessary to execute each action and to deal with the concurrent execution of multiple plans in a timespan. Secondly, we suppose of our agents to be competent in the execution of their plans, i.e. they do not fail to execute their action. This assumption is based on the fact that we are mainly concerned in this work with mental actions and, for that, actual performance is left out of the scope as this would require a formal theory of action phenomenology.

Since only concrete plans can be executed, in the execution of a plan described by the path $\langle \pi, n \rangle$, we must concern ourselves with two cases: $\langle \pi, n \rangle$ is concrete plan, i.e. $n \in leaf(\pi)$, or it is an abstract one.

If it is a concrete plan, the agent will execute it. Since we are assuming the agent is competent in her execution, then the post-condition of the plan will come to hold, i.e. the agent comes to believe with highest certainty that the post-condition comes to hold. To represent this certainty held by the agent, we will update the plausibility of all formulas in the agents - reducing their plausibility - and including the post-condition of the plan in belief base with rank 1.

$$\frac{\mathscr{M} = \langle \pi, n \rangle \in I \quad \text{and} \quad n \in leaf(\pi)}{\langle K, B, G, I \rangle \longrightarrow \langle K, B', G, I' \rangle} \qquad \textbf{(Exec } \pi_1\textbf{)}$$

$$\begin{aligned} \textit{where:} \quad B' &= \{\langle pos(n), 1 \rangle\} \cup \{\langle \varphi, i+1 \rangle \mid \langle \varphi, i \rangle \in B\} \\ I' &= I \setminus \{\langle \pi, n \rangle \in I \mid Bel'(ag) \vDash \neg pre(n) \text{ or } Bel'(ag) \vDash pos(n)\} \end{aligned}$$

If the plan is not yet concrete, the agent must refine it. Notice that plan refinement, as we define here, is an abstraction of the process of selecting between alternative plans to achieve subgoals as commonly done in agent programming languages such as in AgentSpeak (RAO, 1996).

In the rule bellow, we use the notation $I[i/i']$ to express the intention set I with intention i' replacing intention i, i.e. $I[i/i'] = (I \setminus \{i\}) \cup \{i'\}$.

$$\frac{\mathscr{M} = i, \quad i = \langle \pi, n \rangle \in I \quad \text{and} \quad n \notin leaf(\pi)}{\langle K, B, G, I \rangle \longrightarrow \langle K, B, G, I' \rangle} \qquad \textbf{(Exec } \pi_2\textbf{)}$$

$$\begin{array}{lllll} \textit{where:} & i' = \langle \pi, n' \rangle & \text{s.t.} & n' \in suc(n) \text{ and } \langle K, B, G, I[i/i'] \rangle \text{ is coherent.} \\ & I' & = & I[i/i'] \end{array}$$

6.2.2 Acquiring a piece of knowledge

As discussed in Section 6.1, since knowledge is a true belief, it doesn't make sense for an agent to remove a piece of knowledge, only to add one. By the fact that the knowledge base stores all information explicitly known by the agent, it is clear that the act of acquiring a knowledge simply means adding it to the knowledge base. Notice, however, that (i) the transition relation should preserve coherency of agent program states and (ii) for any propositional formula φ and agent ag, $k\varphi$ is an allowed action of ag, as proposed in Definition 6.22. As such, given that coherent agents require that $K \nvDash \bot$, we have to consider the cases in which the acquired knowledge is contradictory with the already established knowledge of the agent and the cases in which it is not.

Let's start with the simplest case. If an agent acquires a knowledge that φ, and φ is consistent with the currently held knowledge, then the resulting state is defined by the inclusion of φ knowledge in the knowledge base. Notice that, by changing the agent's knowledge, the agent may come to discover that the supporting beliefs of some plans, i.e. their pre-conditions, or supporting goals of some plans, i.e. their post-conditions, no longer hold. As such, the agent must drop these intentions, since they are no longer consistent with the agent beliefs. In the rule bellow, we use $Bel'(ag)$ to refer to the beliefs of the agent after change in the agent state and $Goal'(ag)$ to refer to the goals of the agent after change in the agent state.

$$\frac{\mathscr{M} = k\varphi \qquad K \cup \{\varphi\} \nvDash \bot}{\langle K, B, G, I \rangle \longrightarrow \langle K', B, G, I' \rangle} \qquad \textbf{(Know}_1\textbf{)}$$

$$\begin{array}{lll} \textit{where:} \quad K' & = & K \cup \{\varphi\} \\ I' & = & I \setminus (\{\langle \pi, n \rangle \in I \mid Bel'(ag) \vDash \neg pre(n) \text{ or } Bel'(ag) \vDash pos(n)\} \cup \\ & & \{\langle \pi, n \rangle \in I \mid \nexists \xi \in Goal'(ag) \text{ s.t. } pos(n) \vDash \xi\}) \end{array}$$

For the case in which the acquired information φ is inconsistent with the currently held knowledge, we point out once more that the agent knowledge is undefeasible and true, as such the agent always maintains her currently held knowledge. As such, the acquired information

must be wrong, and the agent must discard it.

$$\frac{\mathscr{M} = k\varphi \qquad K \cup \{\varphi\} \vDash \bot}{\langle K,B,G,I\rangle \longrightarrow \langle K,B,G,I\rangle} \qquad (\textbf{Know}_2)$$

6.2.3 Adding a belief

If the agent performs a belief addition action, the agent will simply update her belief base with the new information. Here we must point out that we are accepting AGM's success postulate that the new information is always doxastically more reliable then the currently held beliefs. As such, the newly acquired information will be included with the lowest rank (1) and all the information in the belief base will be made less plausible than it - by increasing the rank of all the items in it. As for the case of change in the knowledge base, by changing the beliefs of the agent, she may come to believe that the supporting beliefs of some plans do not hold and, thus, they need to be abandoned. As such, the rule for addition of a belief is given as follows. As before, we use $Bel'(ag)$ to refer to the beliefs of the agent after change in the agent state.

$$\frac{\mathscr{M} = b^{+}\varphi}{\langle K,B,G,I\rangle \longrightarrow \langle K,B',G,I'\rangle} \qquad (\textbf{Bel}+)$$

$$\begin{aligned} \textit{where:} \quad B' &= \{\langle \varphi, 1\rangle\} \cup \{\langle \psi, i+1\rangle \mid \langle \psi, i\rangle \in B\} \\ I' &= I \setminus \{\langle \pi, n\rangle \in I \mid Bel'(ag) \vDash \neg pre(n) \text{ or } Bel'(ag) \vDash pos(n)\} \end{aligned}$$

Notice that, since the belief base is essentially a stratified base, the simple addition of belief φ with the highest priority is enough to guarantee both success and consistency (consistency is guaranteed by the definition of *maximal consistent subset*, c.f. Definition 6.14)

6.2.4 Removing a belief

If the agent performs a belief removal action, she will contract the belief of her belief base. Notice that the belief base is composed of arbitrary propositional formulas. As such, removing a belief is not simply removing any formula ψ from B that implies φ - since φ can, in general, be implied by subsets of formulas that individually don't imply φ. Some operations have been proposed in the literature for contracting beliefs from stratified belief bases (WILLIAMS, 1994; ROTT, 2009).

Since we aim to interpret the semantic steps by means of the dynamics operations studied in Chapters 4 and 5, we wish that the removal of beliefs work in a way similar to some contraction operation we have studied - particularly, we will use lexicographic contraction. Since the belief base can be understood as a particular case of a priority graph, we can implement such an contraction operation by translating the stratified belief base into a priority graph, computing the contraction and translating the resulting priority graph back to a belief base. Let's first define the priority graph associated with a belief base.

Definition 6.23 *Let $\Gamma \subset \mathscr{L}_0 \times \mathbb{N}$ be a finite set of pairs $\langle \varphi, i \rangle$, called a stratified base. We define the priority graph associated with Γ as $\overline{\Gamma} = \langle \Phi, \prec \rangle$, where $\Phi = \{\bigwedge \Gamma_i$ for all i s.t. $\exists \langle \varphi, i \rangle \in \Gamma\}$ and $\bigwedge \Gamma_i \prec \bigwedge \Gamma_j$ iff $i < j$.*

Since we can translate belief bases into priority graphs, we can apply the lexicographic contraction on such priority graph. What remains is to define the contraction on a stratified base B is to show how to convert a given priority graph into a stratified base. Notice that the priority graph created from a stratified base is linear, i.e. $\prec$ is a total order. Also, the result of the lexicographic contraction will also be a linearly ordered priority graph. As such, we need only to construct a translation of linearly ordered priority graph into a stratified base.

Definition 6.24 *$\mathscr{G} = \langle \Phi, \prec \rangle$ be a finite priority graph, where $\prec$ is a total order. We construct the stratified base $\ulcorner \mathscr{G} \urcorner = \{\langle \varphi, i \rangle \mid \varphi \in \Phi\}$ such that for each $\langle \varphi, i \rangle \in \ulcorner \mathscr{G} \urcorner$, i is the minimal non-null natural number (i.e. $i > 0$) satisfying that for all $\psi \prec \varphi$, $\langle \psi, j \rangle \in \ulcorner \mathscr{G} \urcorner$ with $j < i$.*

Now we can define the lexicographic contraction of a stratified base B by a propositional formula φ as the operation $B -_L \varphi = \ulcorner \overline{B_K} \Downarrow \varphi \urcorner$.

With that, we define the rule for the removal of a belief.

$$\frac{\mathscr{M} = b^{-}\varphi}{\langle K, B, G, I \rangle \longrightarrow \langle K, B', G, I' \rangle} \qquad (\textbf{Bel}-)$$

$$\begin{array}{rrcl} \textit{where:} & B' & = & B -^{L} \varphi \\ & I' & = & I \setminus \{\langle \pi, n \rangle \in I \mid Bel'(ag) \vDash \neg pre(n) \text{ or } Bel'(ag) \vDash pos(n)\} \end{array}$$

6.2.5 Adding a goal

Notice that, different than the encoding of belief, the encoding of goal has an existential nature, in the sense that an agent has a goal to φ if it is consistent with her goal base, instead

of a valid consequence of her goal base. In a way, goals and beliefs have a dual behaviour. As such, to add a new goal φ to the agent goal set, we have to guarantee that φ is consistent with the goal set, i.e. $\neg\varphi$ is not a valid consequence of the maximal consistent subset of the goal base. This can be achieved by contracting the formula $\neg\varphi$ from the agent's goal set.

As before in the removal of a belief, we will use priority graphs to implement the contraction operation we adopt in our semantics in order to guarantee the connection between the semantics of AAP and the logic proposed in Chapter 5. Since we can translate goal bases into priority graphs, as seen in Definition 6.23, we can apply the lexicographic contraction on such priority graph and convert the resulting priority graph to a goal base. Using the notation $G -_L \neg\varphi$ to denote the stratified base $\ulcorner\overline{G} \Downarrow \neg\varphi\urcorner$

With the change in the agents goals, as discussed in Chapter 2, the agent may come to find some alternative more attractive than one of her prior intentions, i.e. the goal ξ supporting a procedural intention $\langle\pi, n\rangle \in I$ may come to be removed from the goal set. Since, to consider an agent state coherent we require that all procedural intentions are supported by a goal, i.e. for all $\langle\pi, n\rangle \in I$ there is a $\xi \in Goal(ag)$, s.t. $pos(n) \vDash \xi$, to maintain mental coherency, we must eliminate all procedural intentions for which the supporting goal is no longer held by the agent. As for the case of beliefs, we will use the notation $Goal'(ag)$ to describe the goals of the agent after updating the goal base. We can, thus, provide the following rule.

$$\frac{\mathscr{M} = g^{+}\varphi}{\langle K, B, G, I\rangle \longrightarrow \langle K, B, G', I'\rangle} \qquad \textbf{(Goal}+\textbf{)}$$

$$\begin{array}{lrcl} \textit{where:} & G' & = & G -^{L} \neg\varphi \\ & I' & = & I \setminus \{\langle\pi, n\rangle \in I \mid \nexists \xi \in Goal'(ag) \text{ s.t. } pos(n) \vDash \xi\} \end{array}$$

6.2.6 Removing a goal

As pointed out in the discussion of goal addition, the agent has a goal to φ if and only if φ is consistent with the maximal consistent subset of her goal base. As such, to remove a goal φ, we have to make sure that $\neg\varphi$ is in the maximal consistent subset of her goal base. This can be guaranteed by introducing $\neg\varphi$ in her goal base with the maximal desirability status.

As before, to maintain mental coherency, the agent must discard all procedural intention for which the supporting goal has been given up by the agent. We use the notation $Goal'(ag)$ to describe the goals of the agent after updating the goal base.

$$\frac{\mathscr{M} = g^{-}\varphi}{\langle K,B,G,I\rangle \longrightarrow \langle K,B,G',I'\rangle} \qquad \textbf{(Goal}-\textbf{)}$$

$$\begin{array}{rrcl} \textit{where:} & G' & = & \{\langle \neg\varphi, 1\rangle\} \cup \{\langle \psi, i+1\rangle \mid \langle \psi, i\rangle \in G\} \\ & I' & = & I \setminus \{\langle \pi, n\rangle \in I \mid \nexists \xi \in \mathit{Goal}'(ag) \text{ s.t. } \mathit{pos}(n) \vDash \xi\} \end{array}$$

6.2.7 Adding an intention

As we have discussed throughout this work, the notion of prospective intentions (or intended state of affairs) implies a set of restrictions on the agent mental state needs to satisfy, encoded by Bratman (1999)'s strong consistency requirements and Cohen and Levesque (1990)'s desiderata. These restrictions are encoded in our language by means of the notion of mental coherency (c.f. Definition 6.21). As such, to adopt a prospective (or declarative) intention φ, we must guarantee that φ satisfies these restrictions, on pain of losing mental state coherency and making our definition unsound.

The first restriction for our notion of prospective intention is that φ is a goal held by the agent ($\varphi \in \mathit{Goal}(ag)$). Second, the agent must believe φ is achievable, i.e. the agent has a plan that can achieve it ($\exists\langle \pi, n\rangle \in \mathscr{H}_{\Pi}$ s.t. $\mathit{pos}(n) \vDash \varphi$) and this plan is pursuable ($ag \vDash B(\mathit{pre}(n))$). More yet, the plan must be consistent with the prior intentions of the agent ($\mathit{pos}(I) \cup \{\mathit{pos}(n)\} \nvdash \bot$), since prior intentions act as a "window of acceptability" for new intentions. A desired state of affairs can, thus, only be adopted as a (declarative) intention of the agent if it satisfy these three requirements. As such, we encode these conditions in the following rule.

$$\frac{\begin{array}{c}\mathscr{M} = i^{+}\varphi \qquad \varphi \in \mathit{Goal}(ag) \qquad \varphi \notin \mathit{Bel}(ag) \\ \exists\langle \pi, n\rangle \in \mathscr{H}_{\Pi} \text{ s.t. } ag \vDash B(\mathit{pre}(n)), \mathit{pos}(n) \vDash \varphi \text{ and } \mathit{pos}(I) \cup \{\mathit{pos}(n)\} \nvdash \bot\end{array}}{\langle K,B,G,I\rangle \longrightarrow \langle K,B,G,I'\rangle} \qquad \textbf{(Int}+_1\textbf{)}$$

$$\textit{where:} \quad I' = I \cup \{\langle \pi, n\rangle\}$$

In the case any of these conditions fails to be satisfied, φ is not admissible to be adopted as an intention.

$$\frac{\begin{array}{c}\mathscr{M} = i^{+}\varphi \text{ and either } \varphi \notin \mathit{Goal}(ag), \varphi \in \mathit{bel}(ag) \text{ or} \\ \nexists\langle \pi, n\rangle \in \mathscr{H}_{\Pi} \text{ s.t. } ag \vDash B(\mathit{pre}(n)), \mathit{pos}(n) \vDash \varphi \text{ and } \mathit{pos}(I) \cup \{\mathit{pos}(n)\} \nvdash \bot\end{array}}{\langle K,B,G,I\rangle \longrightarrow \langle K,B,G,I\rangle} \qquad \textbf{(Int}+_2\textbf{)}$$

6.2.8 Removing an intention

Since intentions are represented as adopted plans, in AAP agent programs, for an agent to give up a declarative intention, she must stop pursuing all the means adopted to achieve it. As such, if the agent gives up her intention to φ, she must stop pursuing any plan that result in φ.

$$\frac{\mathscr{M} = i^{-}\varphi}{\langle K,B,G,I\rangle \longrightarrow \langle K,B,G,I'\rangle} \qquad (\textbf{Int}-)$$

$$\textit{where:} \quad I' \;=\; I \setminus \{\langle \pi, n\rangle \in I \mid pos(n) \vDash \varphi\}$$

6.3 Semantic functions and agent interpreters

Notice that we define a transition relation, not a function. This is because the transition relation $\longrightarrow$ specifies all possible computations, depending on the allowed actions of the agent. We define a concrete interpreter for the language AAP as any function $f : S \to S$ such that $\forall ag \in S : f(ag) = ag' \to \exists a \in Ac_{ag}$ s.t. $\langle ag, a\rangle \longrightarrow ag'$.

To provide a concrete interpreter for the language is, thus, necessary to remove any source of non-determinism in the semantics of AAP. To do this we will use three auxiliary functions.

Notice that the first source of non-determinism in the agent interpretation concerns the decision of which one of the allowed actions to perform at each step of the execution. To select one among all the allowed actions, we will define a step function $\mathscr{M}$. The step function act as a selection tool, guiding the execution of the agent at each step. In the language AgentSpeak (VIEIRA et al., 2007), for example, the step function is implemented by the selection functions.

Definition 6.25 *We call step function, the function $\mathscr{M} : S \to \bigcup_{ag\in S} Ac_{ag}$ such that, for any agent program $ag \in S$, $\mathscr{M}(ag) \in Ac_{ag}$. In other words, for any agent program state ag, the step function selects an allowed action for ag.*

The second source of non-determinism concerns the choice of a plan to achieve a certain goal the agent must select. To deal with this problem, we define the plan selection function S_τ.

Definition 6.26 *Let Π be a set of hierarchical plans. We call plan selection function over Π,*

any function $S_\tau : \mathscr{L}_0 \times S \to \mathscr{H}_\Pi$ *s.t.*

$$S_\tau(\varphi, ag) = \langle \pi, n \rangle \Rightarrow ag \models B(pre(n)), pos(n) \models \varphi \text{ and } pos(I) \cup \{pos(n)\} \not\models \bot$$

Finally, the third source of indetermination in the definition of the transition relation $\longrightarrow$ is in the refinement of a given path or plan in rule Exec π_2. To deal with this problem, we define the path selection function S_ρ.

Definition 6.27 *Let* Π *be a set of hierarchical plans. We call path selection function, any function* $S_\rho : \mathscr{H}_\Pi \times S \to \mathscr{H}_\Pi$ *iff*

$$S_\rho(\langle \pi, n \rangle, ag) = \langle \pi, n' \rangle \Rightarrow ag \models B(pre(n')) \text{ and } n' \in suc(n)$$

Given a step function $\mathscr{M}$ and the selection functions S_τ and S_ρ, we define a concrete interpreter $\longrightarrow_{\mathscr{M},\tau,\rho}$: $S \to S$, where $\forall ag, ag' \in S$ it holds that if $ag \longrightarrow_{\mathscr{M},\tau,\rho} ag'$ then $\langle ag, \mathscr{M}(ag) \rangle \longrightarrow ag'$.

Also, let $ag = \langle K, B, G, I \rangle$ and $\mathscr{M}(ag) = i^+\varphi$. If $\varphi \in Goal(ag)$, $\varphi \notin Bel(ag)$ and there is a plan $\langle \pi, n \rangle$ s.t. $ag \models B(pre(n))$, $pos(n) \models \varphi$ and $pos(I) \cup \{pos(n)\} \not\models \bot$, then $\langle K, B, G, I \rangle \longrightarrow_{\mathscr{M},\tau,\rho} \langle K, B, G, I' \rangle$, and $I' = I \cup \{S_\tau(\varphi, ag)\}$.

If $ag = \langle K, B, G, I \rangle$ and $\mathscr{M}(ag) = \langle \pi, n \rangle \in I$, then $\langle K, B, G, I \rangle \longrightarrow_{\mathscr{M},\tau,\rho} \langle K, B, G, I' \rangle$ and $I' = (I \setminus \{\langle \pi, n \rangle\}) \cup \{S_\rho(\langle \pi, n \rangle, ag)\}$.

Clearly, by definition, for any concrete interpreter f for AAP, there is a step function $\mathscr{M}_f$ and selection functions S_τ and S_ρ such that $f = \longrightarrow_{\mathscr{M}_f,\tau,\rho}$.

6.4 Connecting AAP and Dynamic Preference Logic

As we have presented, the core actions involved in the execution of an agent program may be specified by means of addition and removal of beliefs, goals and intentions of the agent program state.

If we can establish the connection between agent program states, given in Definition 6.8, and agent structures, given in Definition 5.6, we can provide a declarative interpretation of the agent's mental state by means of the correspondence of Theorem 5.7 connecting agent structures and agent models.

Further yet, if we can characterize the rules in the semantic described in Section 6.2 by means of the harmonic operations defined in Chapter 4, the semantics of our programming

language will be subsumed by the declarative specification of the agent.

In this section, we will provide exactly these connections. Let's first establish how a given plan library can be transformed into an plan library as defined in Chapter 5 (Definition 5.11).

Definition 6.28 *Let P be a set of propositional variables and Π a set of hierarchical plans over P. We define the action library $\mathscr{A}_\Pi = \langle A_\Pi, pre_\Pi, pos_\Pi \rangle$ where:*

- $A_\Pi = \{\pi_n$ *a symbol for each node n in a plan $\pi \in \Pi$}*;*
- $pre_\Pi(\pi_n) = pre(n)$*, for $\pi = \langle N, \prec, pre, pos \rangle$ and $n \in N$;*
- $pos_\Pi(\pi_n) = pos(n)$*, for $\pi = \langle N, \prec, pre, pos \rangle$ and $n \in N$;*

Now, we can define how to interpret an AAP agent program over Π as an agent structure for the logic $\mathscr{L}_{\leq_P, \leq_D, \alpha}(P, \mathscr{A}_\Pi)$.

Take an AAP agent program $ag = \langle K, B, G, I \rangle$ over the plan library Π, we want to provide a agent structure $\mathscr{G}_{ag} = \langle \mathscr{G}_P, \mathscr{G}_D \rangle$ that encodes the same information about the knowledge, beliefs, goals and intentions as ag. To do that let's examine how the sets K, B, G and I describe the agent's mental state.

First, the set K, representing the agents knowledge, describes all the information that are undefeasibly true for the agent, as such, any epistemic possible world in a model describing the mental state of ag must satisfy all information in K. From Chapter 5, we know we can guarantee this restriction on the possible worlds performing a public announcement of all the formulas of K in the model representing the agents mental state. Equivalently, we can perform a restriction (as described in Definition 5.18) on the agent structure $\langle \mathscr{G}_P, \mathscr{G}_D \rangle$ representing the agent's beliefs and desires with the formula $\bigwedge K$, the conjunction of all formulas in K, obtaining the agent structure $\langle \mathscr{G}_P^{\bigwedge K}, \mathscr{G}_D^{\bigwedge K} \rangle$.

Second, the stratified bases B, and G describe plausibility and desirability relations among the possible worlds, as such they will define the priority graphs composing the agent structure that represents ag. How to achieve priority graphs $\overline{B}$ and $\overline{G}$ for plausibility and desirability from B and G has been described in Definition 6.23.

Finally, the declarative intentions of an agent describe those desires the agent adopted to achieve. Notice that, since the declarative intentions may act as reasons not to adopt a new goal, they act as overwhelming desires, as discussed in Section 5.2 of Chapter 5. To model this behaviour of declarative intentions as overwhelming desires, we will prefix the priority graph $\overline{G}$ with the singleton graph $\mathscr{G}_I = \langle \{\bigwedge pos(I)\}, \emptyset \rangle$, as described in Definition 4.31, which

corresponds to performing a radical upgrade on the preference relation $\leq_D$ of to agent model representing the agent's mental state. The result of this operation is a priority graph $\mathscr{G}_I;\overline{G}$.

Definition 6.29 *Let $ag = \langle K, B, G, I\rangle$ be an AAP agent program, we define its associated agent structure as $\mathscr{G}_{ag} = \langle \overline{B}^{\wedge K}, (\mathscr{G}_I;\overline{G})^{\wedge K}\rangle$.*

In Definition 6.29, we provide a way to create an agent structure from an AAP agent program. We must guarantee, however that an agent structure constructed in that way represent the same information about the agent's mental state as the information expressed in the AAP program. As such, let's examine the mental attitudes of an agent in both structures.

Proposition 6.30 *Let $ag = \langle K, B, G, I\rangle$ be an AAP agent program over the plan library Π and $\mathscr{G}_{ag}$ its associated agent structure for the logic $\mathscr{L}_{\leq_P,\leq_D}(P, \mathscr{A}_\Pi)$. Then, for any propositional formula $\varphi \in \mathscr{L}_0$, the following holds:*

1. $\varphi \in Know(ag)$ *iff* $\mathscr{G}_{ag} \models A\varphi$
2. $\varphi \in Bel(ag)$ *iff* $\mathscr{G}_{ag} \models B\varphi$
3. $\varphi \in Goal(ag)$ *iff* $\mathscr{G}_{ag} \models G\varphi$
4. *If* $\varphi \in Int(ag)$ *then* $\mathscr{G}_{ag} \models Int\,\varphi$

Proof: The proof is straight-forward from definitions of knowledge, belief, goal and intention in agent programs coupled with Definition 6.29. □

Both Definition 6.29 and Proposition 6.30 establish the connection between AAP agen program states and the priority graphs we named agent structures.

Notice in Proposition 6.30 that the relation between intentions in agent programs and intentions in the logic $\mathscr{L}_{\leq_P,\leq_D}(P, \mathscr{A}_\Pi)$ is not a complete connection, in the sense that it is only a one-way implication. This is because, as we discussed in Chapter 5, intentions in the logic of practical rationality $\mathscr{L}_{\leq_P,\leq_D}(P, \mathscr{A}_\Pi)$ represent which state of affairs the agent **can** consistently intend. To intend something, however, as postulated by Bratman (1999), the agent has to commit and pursue an actual plan to achieve it - which cannot be modelled in the semantics of that logic.

With this connection, we show that, since agent structures are syntactic by nature, they can be understood as data structures that can be used to implement and reason about agent's programs. In fact, an agent program is nothing but a different representation of an agent structure, where the information of knowledge and intention is made explicit.

The connection established is, however, still limited to the static behaviour of program states, i.e. not to how these states evolve given the language semantics provided in Section 6.5. Our aim now is to connect the semantic rules describing the transition relation $\longrightarrow$ with the harmonic operations described in Chapter 5.

In investigating the declarative interpretation for the execution of an AAP agent program, it will be useful to speak not of the agent's agent structure but of the agent's mental state, i.e. an agent model describing the agent's mental state. Since we can establish a connection between agent structures and agent models, we can provide such a model without difficulty. In the Definition 6.31 below, we take the mental state of an AAP agent *ag* as a broad model induced by $\mathscr{G}_{ag}$ after the public announcement of $\bigwedge K$.

Definition 6.31 *Let $ag = \langle K, B, G, I\rangle$ be an AAP agent program over the plan library Π and $\mathscr{G}_{ag}$ its associated agent structure. We define the model $M_{ag} = \langle W, \leq_P, \leq_D, v\rangle$, called the mental state of agent ag, where:*

- $W = \{w \in 2^P \mid w \vDash_{\mathscr{L}_0} \bigwedge K\}$;
- $\leq_P = \leq_{\overline{B}}$;
- $\leq_D = \leq_{(\mathscr{G}_I;\overline{G})}$;
- *for any $p \in P$, $w \in v(p)$ iff $p \in w$.*

With that, we can show that the evolution of the agent program can be completely described by the changes occurring in the agent's mental state. In other words, that we can describe the operational semantics of AAP by the declarative semantics provided by the logic $\mathscr{L}_{\leq_P,\leq_D}(P, \mathscr{A}_\Pi)$. As such, this logic can be used to perfectly describe the behaviour of AAP programs.

Let's start with the simple case of acquiring knowledge in agent programs. It is clear that from the definition of knowledge in agent programs and from the construction of the associated agent structure that the change described by acquiring a new knowledge φ - not inconsistent with the currently held knowledge - corresponds to guarantee that any world in the model describing the agent's mental state satisfies this information. This is exactly the result of performing a public announcement of φ in the model representing the agent's mental state as described in Definition 5.17, in the agent structure describing this model.

Proposition 6.32 *Let $ag = \langle K, B, G, I\rangle$ and $ag' = \langle K', B, G, I'\rangle$ be AAP agent states and let $\langle ag, k\varphi\rangle \longrightarrow ag'$, for some propositional formula φ s.t. $K \nvDash \neg\varphi$. Then $M_{ag'} = (M_{ag})_{!\varphi}$, the public announcement φ in M_{ag}.*

Now let's examine more interesting changes in the agent's beliefs and goals and intentions. Since changes in the agent's beliefs and goals imply a change in her intentions, let's begin investigating the change in her intentions.

When an agent adopts a procedural intention $\langle \pi, n \rangle$ to achieve a goal φ, by construction of the graph $\mathscr{G}_a p$ in Definition 6.29, it corresponds to her promoting the goal $pos(n)$. As such, $pos(n)$ will become a overwhelming desire in her mental state. To make a goal $pos(n)$ the most desirable in the agents mental state corresponds, thus, to perform a desirability update by radical upgrade of the formula $pos(n)$ in the mental state of ag.

Proposition 6.33 *Let $ag = \langle K, B, G, I \rangle$ and $ag' = \langle K, B, G, I' \rangle$ be AAP agent states and let $\langle ag, i^+\varphi \rangle \longrightarrow ag'$ s.t. $\varphi \in Goal(ag)$, for some propositional formula φ. If $I' = I \cup \{\langle \pi, n \rangle\}$, i.e. $ag \vDash B(pre(n))$, $pos(n) \vDash \varphi$, $pos(n) \notin Bel(ag)$ and $pos(I) \cup \{pos(n)\} \nvdash \bot$, then $M_{ag'} = (M_{ag})_{\Uparrow_D pos(n)}$, where $(M_{ag})_{\Uparrow_D pos(n)}$ is the desirability upgrade of M_{ag} by radical upgrade of $pos(n)$ as presented in Definition 5.22.*

To remove an intention to φ corresponds to remove all procedural intentions $\langle \pi, n \rangle$ to achieve φ. Notice, however, in Definition 6.29, that the post-conditions of I (denoted by $pos(I)$) are taken as the most desirable in the agent structure $\mathscr{G}_{ag}$. The effect of removing from I all plans satisfying φ, resulting in the set I', corresponds then to elevate the desirability of the worlds of M_{ag} in which $\bigwedge pos(I') \wedge \neg\varphi$ hold. This is achieved by making the worlds satisfying $\bigwedge pos(I') \rightarrow \varphi$ less desirable, i.e. performing a contraction of $\bigwedge pos(I') \rightarrow \varphi$ in M_{ag},

Notice also that intentions are overwhelming desires. This restriction is made explicit in Definition 6.29 by prefixing $\overline{G}$ by the graph $\mathscr{G}_I = \langle \{\bigwedge pos(I)\}, \emptyset \rangle$. Thus the goals represented in $pos(I)$ are maximal in the model representing the mental state of an agent program. This, however, implies that some information may be redundant in the resulting agent structure - namely, all of the agent's overwhelming desires which were adopted as intentions are doubly represented in the graph $pos(I); \overline{G}$. To make it more clear, let's examine Example 6.34

Example 6.34 *Take the agent $ag = \langle \emptyset, \emptyset, \{(a,1),(b,2)\}, \{\langle \pi, n \rangle\} \rangle$ where $\pi = \langle \{n\}, pre, pos, \emptyset \rangle$, $pre(n) = \top$ and $pos(n) = a$. By Definition 6.23, $\overline{G} = \langle \{a,b\}, \{\langle a,b \rangle\} \rangle$ and $\mathscr{G}_I; \overline{G} = \langle \{a,b\}, \{\langle a,b \rangle\} \rangle$. As such, if the agent removes the intention to a, we obtain the program $ag' = \langle \emptyset, \emptyset, \{(a,1),(b,2)\}, \emptyset \rangle$, where $\mathscr{G}_\emptyset; \overline{G'} = \overline{G}$.*

As shown in Example 6.34, however, we cannot differentiate in the graph $\mathscr{G}_I; \overline{G'}$ if the agent has an overwhelming desire to φ due to her goals or her intentions. As such, to study the semantics of the programming language by means of the declarative interpretation, we will

consider only the changes which are performed in the agent's mental state when the intended goal is not an overwhelming goal in the agent's goal base.

Proposition 6.35 *Let $ag = \langle K,B,G,I\rangle$ and $ag' = \langle K,B,G,I'\rangle$ be AAP agent states and let $\langle ag, i^{-}\varphi\rangle \longrightarrow ag'$ and $\neg\varphi \in Goal(ag)$, for some propositional formula φ. Then $M_{ag'} = (M_{ag})_{\Downarrow_D \bigwedge pos(I')\rightarrow\varphi}$, where $(M_{ag})_{\Downarrow_D \bigwedge pos(I')\rightarrow\varphi}$ is the desirability update of M_{ag} by contraction of $pos(I') \rightarrow \varphi$ as presented in Definition 5.26.*

In the case in which the intended goal is a overwhelming desire, the operations of abandoning a plan has no effect in the agent's mental state, as long as the logic $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$ is concerned. The reason for that is that in this case the mental operation is restricted to what plans the agent is *actually* pursuing, which cannot be represented in that logic.

To add a belief φ in an agent program ag, we give the formula φ the highest plausibility on her base, i.e. we guarantee that φ is believed by the agent. This is similar to what we do in the transformation of graph prefixing presented in Definition 4.31. In this way, we point out that the belief addition in AAP behaves similarly as the operation of radical upgrade. As the intentions may change after such change, we need to update her desirability relation as well.

Proposition 6.36 *Let $ag = \langle K,B,G,I\rangle$ and $ag' = \langle K,B',G,I'\rangle$ be AAP agent states and let $\langle ag, b^{+}\varphi\rangle \longrightarrow ag'$, for some propositional formula φ. Then $M_{ag'} = ((M_{ag})_{\Uparrow_P\varphi})_{\Downarrow_D pos(I')\rightarrow\varphi}$, where $(M_{ag})_{\Uparrow_P\varphi}$ is the plausibility update of M_{ag} by radical upgrade of φ as presented in Definition 5.20 .*

Since we implemented removal of beliefs by means of lexicographic contraction on a priority graph, in rule $Bel-$, it is not difficult to see that, by construction, it corresponds to the plausibility update by lexicographic contraction.

Proposition 6.37 *Let $ag = \langle K,B,G,I\rangle$ and $ag' = \langle K,B',G,I'\rangle$ be AAP agent states and let $\langle ag, b^{-}\varphi\rangle \longrightarrow ag'$, for some propositional formula φ. Then $M_{ag'} = ((M_{ag})_{\Downarrow_P\varphi})_{\Downarrow_D pos(I')\rightarrow\varphi}$. where $(M_{ag})_{\Downarrow_P\varphi}$ is the plausibility update of M_{ag} by contraction of φ as presented in Definition 5.26.*

As we have seen in Section 6.2, goal addition and removal, and belief addition and removal have dual behaviours, in the sense that one add a goal by contraction and removes a goal by addition. This is because beliefs and goals are defined by very different means, beliefs are universally quantified, goals are existential quantified. Based on the two results above, the following is thus not surprising.

Proposition 6.38 *Let $ag = \langle K,B,G,I\rangle$ and $ag' = \langle K,B,G',I'\rangle$ be AAP agent states and let $\langle ag, g^{+}\varphi\rangle \longrightarrow ag'$, for some propositional formula φ. Then $M_{ag'} = ((M_{ag})_{\Downarrow_D \neg\varphi})_{\Downarrow_D pos(I')\rightarrow\varphi}$, where $(M_{ag})_{\Downarrow_D\neg\varphi}$ is the desirability update of M_{ag} by contraction of $\neg\varphi$ as presented in Definition 5.26.*

Also, since goal removal is performed similarly as belief addition, we have the following result.

Proposition 6.39 *Let $ag = \langle K,B,G,I\rangle$ and $ag' = \langle K,B,G',I'\rangle$ be AAP agent states and let $\langle ag, g^{-}\varphi\rangle \longrightarrow ag'$, for some propositional formula φ. Then $M_{ag'} = ((M_{ag})_{\Uparrow_D \neg\varphi})_{\Downarrow_D pos(I')\rightarrow\varphi}$, where $(M_{ag})_{\Uparrow_D\neg\varphi})$ is the desirability update of M_{ag} by radical upgrade of $\neg\varphi$ as presented in Definition 5.22.*

The last rule to be investigated is the execution of a plan. We know from Proposition 5.16 that after the execution of a plan, the agent believes in its post-conditions. As such, it is clear that the rule *Exec* π_1 is safe, in the sense that it preserves this behaviour. The problem, however, is that we cannot understand the execution of a plan in AAP as a mental attitude in DLP, as with the other cases.

If the plan is concrete, in fact, due to our requirement that the agent is competent in the execution, it amounts to perform an addition of the belief that the post-condition of the plan comes to hold.

Proposition 6.40 *Let $ag = \langle K,B,G,I\rangle$ and $ag' = \langle K,B',G,I'\rangle$ be AAP agent states and let $\langle ag, i\rangle \longrightarrow ag'$, with $i = \langle\pi,n\rangle \in I$ and $n \in leaf(\pi)$. Then $M_{ag'} = ((M_{ag})_{\Uparrow_P pos(n)})_{\Downarrow_D pos(I')\rightarrow\varphi}$, where $(M_{ag})_{\Uparrow_P pos(n)}$ is the plausibility update of M_{ag} by radical upgrade of $pos(n)$ as presented in Definition 5.20.*

If the plan is abstract, however, we must refine it, which amounts to adopt a new intention.

Proposition 6.41 *Let $ag = \langle K,B,G,I\rangle$ and $ag' = \langle K,B,G,I'\rangle$ be agent states s.t. $\langle ag, i\rangle \longrightarrow ag'$, $i = \langle\pi,n\rangle \in I$ and $n \notin leaf(n)$. If $I' = I[i/i']$ and $i' = \langle\pi,n'\rangle$, then $M_{ag'} = (M_{ag})_{\Uparrow_D pos(n')}$, where $(M_{ag})_{\Uparrow_D pos(n')}$ is the desirability update of M_{ag} by radical upgrade of $pos(n)$ as presented in Definition 5.22.*

Notice that, while we investigate the semantics of our very simple abstract language AAP, the semantics of many agent programming languages are defined by the manipulation of

information structures similar to what we call an agent program. In fact, it is not uncommon that the semantics of BDI languages is defined by a transition relation in a similar way to what we presented, e.g. we urge the interested reader to see the specification of the AgentSpeak (VIEIRA et al., 2007) and GOAL (HINDRIKS et al., 2001) programming languages. We claim that the mental operations discussed here are enough to model the reasoning cycle of agents in most agent programming languages.

6.5 There and back again: returning to the philosophical considerations

Once presented the logic we use to encode mental states and the methodology we propose to study the semantics of agent programming languages by means of this logic, we return to compare our proposal to the philosophical foundations we claim for our work.

First let's examine the meaning of intentions in our framework. It is evident that in our work we subscribe to Bratman (1999)'s intentions as plans view. We do this firstly because the aim of our work is to study the semantics of agent programming language and, this is the most predominant view for their foundation. Second, and perhaps most important, we believe this view is pragmatically adequate for agent programs: giving the semantic foundation of practical reasoning a convenient computational model.

As patent in the encodings in both our logic and in the semantics of the abstract agent programming language AAP, we require that intention satisfy Bratman's strong consistency requirements. As for our encoding in the logic, it is not difficult to see from Proposition 6.30 that our computational encodings of mental attitudes satisfy Cohen and Levesque (1990)'s desiderata for intentions, as well.

In accepting these requirements for the notion of intention, we accept that unfortunate side-effect of disallowing intentions that otherwise would be natural for an agent to hold - as in the aforementioned Video Game Puzzle - in order to keep the Simple View which states that to perform an action intentionally, the agent must have an intention to perform this action. This view, as discussed in Chapter 2, is inconsistent with Bratman (1999)'s strong consistency requirements since it is conceivable in some cases to hold intentions that are inconsistent with each other. An example of such a case is argued by Bratman in the Video Game Puzzle.

Our response to this criticism is to disallow the agent to hold the mutually inconsistent intentions at the same time, i.e. the agent must choose beforehand which of the intentions she wishes to pursue. As such, as it has been done in the formal literature on intentions and in their reproduction in computational system, we accept the Simple View as a tool to unify the notions

of prospective intention and intentional action - going against Bratman's conclusions. We claim that these choices are justifiable in the context of Agent Programming, but we do not claim that it can be extended further to explain human (or animal) agency in Cognitive Science (Ethology) without its problems.

Regarding the problem of intention reconsideration, we notice that our encoding supports Mintoff (2004)'s principles for intention reconsideration discussed in Chapter 2. In other words, an intention is reconsidered if, and only if, either the beliefs supporting this intention are reconsidered or there is a change in the agent's desires that provide reasons for the agent to reconsider her intention. This is a felicitous side-effect of the encodings and the semantics of change in DPL, since we did not consciously pursued this connection. Notice that, in our semantics, an agent comes to reconsider and intention iff she no longer believes it to be possible to achieve it or if the goal associated with the intention is no longer a preferred goal, i.e. the agent comes to find there is a possibly more attractive alternative.

6.6 Summary of the chapter

In this chapter, we propose an abstract agent programming language AAP and show in Section 6.4 that an agent program in this language can be understood as an agent structure presented in Chapter 5. In Section 6.1 we define the notion of an agent program and the encodings of mental attitudes in it. In Section 6.2, we define the operational semantics of the language AAP, given by a set of transition rules. In Section 6.3, we discuss how the semantics provided in Section 6.2 can be implemented into a concrete interpreter - by removing all non-determinism in the transition rules. In Section 6.4, we connect the semantics of AAP with that of the language $\mathscr{L}_{\leq_P,\leq_D}(P\mathscr{A})$ presented in Chapter 5, by means of interpreting an agent program as an equivalent agent structure. Finally, in Section 6.5, we revisit the discussions presented in Chapter 2 and whether the language AAP satisfy the principles we adopted.

7 FINAL CONSIDERATIONS

This work has studied the semantics of mental attitudes and its dynamics from the perspective of Agent Programming. With a strong focus on the philosophical foundations, we proposed a logic for practical rationality and proposed a connection between this logic and agent programming languages by means of Liu (2011)'s representation theorems for preference models.

The main difference of our work from that of the literature is that we propose to study the dynamics of mental attitudes as a guiding theme both for their formal representation in the logic, as well as for its implementation in an agent programming language. Other studies in the literature have focused either on the static requirements for mental coherence, i.e. the relationship between the agent's mental attitudes, or on their change as an effect of the actions of the agent into her environment and, as such, a result of agent's perception.

We focus, in this thesis, on mental actions instead of ontic actions, i.e. actions that affect not the environment in which the agent is embodied, but the agent's mind. We believe mental actions are an adequate tool to describe and analyse the cognitive steps involved in the problem of practical rationality. By this focus, we believe our study of mental attitude dynamics to be much more relatable with the inner workings of actual agent programming languages and, as such, more applicable to the study of their semantics. This is because most agent programming languages limit themselves in describing the cognitive aspects of practical rationality, i.e. the mental mechanisms involved in selection of action, not on actual performance of action.

By interpreting stratified bases commonly used in agent programming languages to describe an agent's mental state, we provide a general way to achieve a declarative interpretation of agent programs - by means of the connection between agent structures and agent models described in Chapter 5. Also, giving that the semantics of these languages are defined by means of consecutive changes in the agent's mental state - concerning actions such as adopting a goal, a plan, a belief, etc. - our dynamic logic can be used to faithfully study the change in agent's mental attitudes as governed by the programming language semantics.

Regarding the specific goals of our thesis, as listed in Section 1.1 in the Introduction, we notice that:

- The first goal, i.e. to provide a logic that encodes both static and dynamic requirements of mental attitudes as discussed in the philosophical work in BDI theory, has been achieved by the dynamic logic of rationality and practicality proposed in Chapter 5;
- The second goal, i.e. to provide a set of 'safe' operations on the models that can be used to

characterize the mental actions involved in practical reasoning, has also been achieved in Chapter 5 with the operations of public announcement, radical upgrade, public suggestion and lexicographic contraction;

- The third goal, i.e. to characterize these operations by means of syntactic operations over priority graphs, has been achieved in Chapters 4 and Chapter 5;
- The fourth goal, i.e. to study the relation between declarative agent specification logics and logic programming semantics, has been achieved Chapter 6 for an abstract agent programming language.

7.1 Results of our work

A first result of our studies concerns some general results on Girard (2008)'s Dynamic Logic of Preferences, presented in Chapter 4. With our work, we proved a complete axiomatization for the DLP defined over well-founded preference models, validating Girard and Rott (2014)'s conjecture that the Löb Axiom (W) for the strict accessibility relation $<$ is enough to guarantee well-foundedness of the model.

Additionally, we provided semantic codifications for three different contraction operations in this logic and the representation for them by means of priority graphs. Also, by studying contraction operations in this setting, we provided an answer to whether all PDL-definable operations that preserve preference models are representable as transformation on priority graphs.

A second immediate result of our work is our logic of practical rationality, proposed in Chapter 5. In this logic, we encode notions of BDI mental attitudes and provide semantic characterizations of some change operations on an agent's mental state. Also, by using Liu (2011)'s representation results for DPL, we characterize an agent mental state and its dynamics by means of agent structures and safe operations over these structures.

Finally, to study how to define a declarative (or denotational) semantics for agent programming languages, we propose an abstract agent programming language in Chapter 6, called AAP, that contain some of the most basic aspects of an agent programming language. We provide a formal semantics for this language and show how to connect the formal semantics of the language with the declarative semantics provided by the logic of practical rationality. We show that the actions defined in this programming language can be understood as the semantic operations defined in DPL and, by that, show that the formal semantics of AAP can be understood by means of transformations on the declarative interpretation of an agent's mental state.

7.2 Publications

On studying the semantics of communication between agents and its effects on agent's mental states, we have provided some semantic conditions for effective communication for heterogeneous agent's with ontological reasoning and proposed a abstract communication mechanism satisfying these conditions. The results of our work have been published in:

- SOUZA, M.; MOREIRA, A.; VIEIRA, R; MEYER, J.J. C. Communication for Agents with Ontological Reasoning. In: **Proceedings of the 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)**. v. 2, p. 182–185. 2015.
- SOUZA, M.; MOREIRA, A.; VIEIRA, R; MEYER, J.J. C. Integrating Ontology Negotiation and Agent Communication. **Ontology Engineering: 12th International Experiences and Directions Workshop on OWL, OWLED 2015, co-located with ISWC 2015**. p. 56–68. Springer. 2016.

Also, from our study about harmonic operation on DPL, i.e. operations that can be equivalently defined by means of PDL transformations on preference models or syntactic transformations on priority graphs, we published the following paper:

- SOUZA, M.; MOREIRA, A.; VIEIRA, R; MEYER, J.J. C. Preference and priorities: a study based on contraction. In: **Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning (KR)**. p. 155–164. AAAI Press. 2016.

From our studies in the formal semantics of a BDI language with underlying ontological reasoning including doxastic change and agent communication, we submitted the paper:

- SOUZA, M.; MOREIRA, A.; VIEIRA, R. Ontology Reasoning in Agent-Oriented Programming. Artificial Intelligence Review. 37 p.

7.3 Future directions

Regarding future developments of our work, we expect to achieve contributions in two different areas: agent programming semantics and belief change theory.

From the point of view of agent programming, the first contribution is to apply our abstract language AAP to study the semantics of a concrete agent programming language, i.e.

a language proposed in the literature and with a working interpreter. This work has, in fact, been advanced by us and a translation of a programs following the PRS-like architecture, a generalization of the PRS architecture proposed by Wobcke (2001), and AAP programs has been studied by us. The connections between these two languages is not yet fully understood, since the PRS-like does not possess any notion of declarative goals.

Second, we expect to use our language of rationality to study the semantics of agent programming languages extended with theoretical reasoning rules, such as the integration of ontologies in AgentSpeak-DL (MOREIRA et al., 2006). We believe this extended framework to be a fruitful environment to study the relationship of beliefs, desires and intentions an their dynamics in programming languages.

Although no formal connection between AAP and AgentSpeak-DL has been provided yet, our study has highlighted several mechanisms in the semantics of the latter, which were not well understood from a theoretical perspective. For example, while AgentSpeak and most agent languages based on the PRS architecture have only a procedural notion of intention, AgentSpeak-DL employs a declarative interpretation of agents' goals while using the theoretical reasoning rules, based on the interpretation provided by Bordini and Moreira (2004).

As such, one can say that a certain theoretical rule $A \sqsubseteq B$ can act as a reason for an agent to adopt a plan $!A(t_0)$ when trying to achieve a goal $B(t_0)$. If the agent comes to no longer believe $A \sqsubseteq B$, however, this rule cannot be held as a reason for the agent to achieve $!A(t_0)$. This, and some other mechanism, have been identified while contrasting the semantics of AgentSpeak-DL and our framework, leading to the proposal of a revised semantics for the language - submitted for publication. A formal connection between our framework and the semantics of the language must yet be attempted.

From point of view of our formal theory, we also wish to expand our foundational logic to model action execution and action attempt, such as in the logic of Wobcke (2004). This extension would allow us to bridge the semantic difference of intentions in agent programming languages - as currently held plans - and intentions in the specification language - as possibility of action. This would allow us cross between agent specifications in the logic and agent executions in the programming language in a completely transparent manner, i.e. without any loss of information in either side.

On this last topic, we point out that the integration of the cognitive aspects of agency and decision making and its phenomenal aspects, such as perceptions and action execution, is a topic that has been studied in the literature, e.g. by van Oijen, Vanhée and Dignum (2012). As such, we can integrate their approach in our framework to provide a wholesome view of the

agent's behaviour in a situated manner.

Recently, Herzig et al. (2016) pointed out some deficiencies in the formal frameworks for specifying BDI agents which are available in the literature. Similar to our criticisms, these authors point out the advantages of a formal theory with a close relationship with the work in belief dynamics and with agent programming - also indicating Dynamic Epistemic Logic as a promising framework to base such a theory.

We believe our work tackles most requirements these authors list for a formal theory of agent programming. It remains, however, to provide a greater connection of our logics with the work areas as *planning* and *game theory*. We point out, however, that we have powerful evidences that such connections can be done. For the connection with planning, the work of Andersen, Bolander and Jensen (2014) explore the semantics of dynamic epistemic logic with complex ontic actions, i.e. how to integrate planning in the dynamic epistemic logics as the one we propose. For the connection with game theory, the work of Roy (2009) provides a dynamic epistemic logic for intentions with a semantics based on epistemic game theory. We believe we can provide a connection between our agent models and the possible world semantics based on game strategies, as proposed by the author in his work.

From the point of view of belief change theory, we believe our work indicates that Dynamic Preference Logic is a powerful framework able to provide a comprehensive model theory for the logic of belief change. This theory can faithfully absorb and generalise the techniques employed in the area of Iterated Belief Revision and is able to represent the difference in expressibility of operations performed on the syntactic level, such as in belief base change (HANSSON, 1992), by priority graphs, and on the semantic level, such as belief revision (ALCHOURRÓN; GÄRDENFORS; MAKINSON, 1985), by preference models.

Also we point out that, based on the work of Zhang and Ding (2008), since we can determine an ordering among CTL models based on a specification, we believe DPL is able to represent the (more general) operations of model change proposed by those authors.

Using the connection between Kripke Modal Transition Systems (HUTH; JAGADEESAN; SCHMIDT, 2001) and CTL models (EMERSON; HALPERN, 1985), Guerra, Andrade and Wassermann (2013) show how to compute CTL model revision by means of revisions in a KMTS, in similar way to defining operations on models by means of priority graphs. As such, we believe we may employ the techniques discussed in Section 4 to investigate model revision, using KMTS as a graphic representation of a preferential model, in which each possible world is a CTL model and the preference relation is established by Zhang and Ding (2008)'s ordering.

REFERENCES

ALCHOURRÓN, C. E.; GÄRDENFORS, P.; MAKINSON, D. On the logic of theory change: Partial meet contraction and revision functions. **Journal of Symbolic Logic**, v. 50, n. 2, p. 510–530, 1985.

ALCHOURRÓN, C. E.; MAKINSON, D. On the logic of theory change: Contraction functions and their associated revision functions. **Theoria**, John Wiley & Sons, Hoboken, US, v. 48, n. 1, p. 14–37, 1982.

ALECHINA, N. et al. Belief revision for AgentSpeak agents. In: **Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems**. New York, US: ACM, 2006. p. 1288–1290.

ALECHINA, N. et al. Reasoning about agent deliberation. **Autonomous Agents and Multi-Agent Systems**, Springer, New York, US, v. 22, n. 2, p. 356–381, 2011.

ANDERSEN, M. B.; BOLANDER, T.; JENSEN, M. H. Don't plan for the unexpected: Planning based on plausibility models. **Logique et Analyse**, v. 1, n. 1, 2014.

ANDRÉKA, H.; RYAN, M.; SCHOBBENS, P.-Y. Operators and laws for combining preference relations. **Journal of logic and computation**, Oxford University Press, Oxford, UK, v. 12, n. 1, p. 13–53, 2002.

ANSCOMBE, G. E. M. **Intention**. Cambridge, US: Harvard University Press, 1957.

AUDI, R. Intending. **The Journal of Philosophy**, JSTOR, p. 387–403, 1973.

BALTAG, A.; FIUTEK, V.; SMETS, S. DDL as an "internalization" of dynamic belief revision. In: **Krister Segerberg on Logic of Actions**. Dordrecht, NL: Springer Netherlands, 2014. p. 253–280.

BALTAG, A.; MOSS, L. S.; SOLECKI, S. The logic of public announcements, common knowledge, and private suspicions. In: **Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge**. Burlington, US: Morgan Kaufmann, 1998. p. 43–56.

BALTAG, A.; SMETS, S. A qualitative theory of dynamic interactive belief revision. **Texts in logic and games**, Amsterdam University Press, v. 3, p. 9–58, 2008.

BENFERHAT, S. et al. Inconsistency management and prioritized syntax-based entailment. In: **Proceedings of the 13th International Joint Conference on Artificial Intelligence**. Burlington, US: Morgan Kaufmann, 1993. v. 93, p. 640–645.

BLACKBURN, P.; VAN BENTHEM, J. F.; WOLTER, F. **Handbook of modal logic**. Amsterdam, NL: Elsevier, 2006.

BORDINI, R.; HUBNER, J.; WOOLDRIDGE, M. **Programming Multi-agent Systems in AgentSpeak Using Jason**. Hoboken, US: John Wiley & Sons, 2007.

BORDINI, R.; MOREIRA, A. Proving BDI properties of agent-oriented programming languages: The asymmetry thesis principles in AgentSpeak (L). **Annals of Mathematics and Artificial Intelligence**, Springer, New York, US, v. 42, n. 1, p. 197–226, 2004.

BOUTILIER, C. Revision sequences and nested conditionals. In: **Proceedings of the 13th International Joint Conference on Artificial Intelligence**. New York, US: Morgan Kaufmann, 1993. v. 93, p. 519–531.

BOUTILIER, C. Conditional logics of normality: a modal approach. **Artificial Intelligence**, Elsevier, Amsterdam, NL, v. 68, n. 1, p. 87–154, 1994.

BOUTILIER, C. Toward a logic for qualitative decision theory. In: **Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning**. New York, US: Morgan Kaufmann, 1994. p. 75–86.

BRATMAN, M. E. Planning and the stability of intention. **Minds and Machines**, Springer, New York, US, v. 2, n. 1, p. 1–16, 1992.

BRATMAN, M. E. **Intention, plans, and practical reason**. Cambridge, US: Harvard University Press, 1999.

BRATMAN, M. E.; ISRAEL, D. J.; POLLACK, M. E. Plans and resource-bounded practical reasoning. **Computational intelligence**, John Wiley & Sons, Hoboken, US, v. 4, n. 3, p. 349–355, 1988.

BROERSEN, J. et al. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In: ACM. **Proceedings of the fifth international conference on Autonomous agents**. New York, US, 2001. p. 9–16.

BROERSEN, J. et al. Goal generation in the BOID architecture. **Cognitive Science Quarterly**, v. 2, n. 3-4, p. 428–447, 2002.

CASTELFRANCHI, C. Mind in degrees. In: **XIIIth International Conference of the Italian Association for Artificial Intelligence**. New York, US: Springer, 2013. p. 13–24.

CASTELFRANCHI, C.; PAGLIERI, F. The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. **Synthese**, Springer, New York, US, v. 155, n. 2, p. 237–263, 2007.

CHELLAS, B. F. **Modal logic: an introduction**. Cambridge, UK: Cambridge University Press, 1980.

CHISHOLM, R. M. Contrary-to-duty imperatives and deontic logic. **Analysis**, JSTOR, v. 24, n. 2, p. 33–36, 1963.

CHURCHLAND, P. M.; CHURCHLAND, P. S. Functionalism, qualia, and intentionality. **Philosophical Topics**, JSTOR, v. 12, n. 1, p. 121–145, 1981.

COHEN, P. R.; LEVESQUE, H. J. Intention is choice with commitment. **Artificial Intelligence**, v. 42, n. 2-3, p. 213–261, 1990.

DARWICHE, A.; PEARL, J. On the logic of iterated belief revision. **Artificial intelligence**, Elsevier, Amsterdam, NL, v. 89, n. 1, p. 1–29, 1997.

DASTANI, M. 2APL: a practical agent programming language. **Autonomous agents and multi-agent systems**, Springer, New York, US, v. 16, n. 3, p. 214–248, 2008.

DASTANI, M. et al. Debugging BDI-based multi-agent programs. In: **Programming Multi-Agent Systems**. Berlin, DE: Springer-Verlag, 2010. p. 151–169.

DASTANI, M.; HULSTIJN, J.; VAN DER TORRE, L. BDI and QDT: a comparison based on classical decision theory. In: **Game Theoretic and Decision Theoretic Agents**. Palo Alto, US: AAAI Press, 2001. p. 16–26.

DASTANI, M. et al. A programming language for cognitive agents goal directed 3apl. In: **Programming Multi-Agent Systems**. New York, US: Springer, 2003. p. 111–130.

DASTANI, M.; VAN RIEMSDIJK, B.; MEYER, J.-J. C. Programming multi-agent systems in 3apl. In: **Multi-agent programming**. New York, US: Springer, 2005. p. 39–67.

DASTANI, M.; VAN RIEMSDIJK, M. B.; MEYER, J.-J. C. A grounded specification language for agent programs. In: **Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems**. New York, US: ACM, 2007. p. 147.

DAVIDSON, D. Intending. In: **Philosophy of history and action**. New York, US: Springer, 1979. p. 41–60.

DE BOER, F. et al. A verification framework for agent programming with declarative goals. **Journal of Applied Logic**, Elsevier, Amsterdam, NL, v. 5, n. 2, p. 277–302, 2007.

D'INVERNO, M. et al. A formal specification of dmars. In: **Intelligent Agents IV Agent Theories, Architectures, and Languages**. New York, US: Springer, 1998. p. 155–176.

DOYLE, J. A truth maintenance system. **Artificial intelligence**, Elsevier, Amsterdam, NL, v. 12, n. 3, p. 231–272, 1979.

DOYLE, J.; SHOHAM, Y.; WELLMAN, M. P. A logic of relative desire. In: **Methodologies for Intelligent Systems**. New York, US: Springer, 1991. p. 16–31.

EMERSON, E. A.; HALPERN, J. Y. Decision procedures and expressiveness in the temporal logic of branching time. **Journal of computer and system sciences**, Elsevier, Amsterdam, NL, v. 30, n. 1, p. 1–24, 1985.

FERMÉ, E.; KREVNERIS, M.; REIS, M. An axiomatic characterization of ensconcement-based contraction. **Journal of Logic and Computation**, Oxford University Press, Oxford, UK, v. 18, n. 5, p. 739–753, 2008.

FRANKLIN, S.; GRAESSER, A. Is it an agent, or just a program?: A taxonomy for autonomous agents. In: **Intelligent agents III agent theories, architectures, and languages**. New York, US: Springer, 1997. p. 21–35.

GARAU, M. et al. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In: **Proceedings of the SIGCHI conference on Human factors in computing systems**. New York, US: ACM, 2003. p. 529–536.

GÄRDENFORS, P.; MAKINSON, D. Revisions of knowledge systems using epistemic entrenchment. In: **Proceedings of the 2nd conference on Theoretical aspects of reasoning about knowledge**. Burlington, US: Morgan Kaufmann, 1988. p. 83–95.

GEORGEFF, M. P.; LANSKY, A. L. Reactive reasoning and planning. In: **Proceedings of the Sixth National Conference on Artificial Intelligence**. Palo Alto, US: AAAI Press, 1987. p. 677–682.

GIERASIMCZUK, N. Bridging learning theory and dynamic epistemic logic. **Synthese**, Springer, New York, US, v. 169, n. 2, p. 371–384, 2009.

GIRARD, P. **Modal logic for belief and preference change**. Thesis (PhD) — Stanford University, 2008.

GIRARD, P.; ROTT, H. Belief revision and dynamic logic. In: **Johan Van Benthem on Logic and Information Dynamics**. New York, US: Springer, 2014. p. 203–233.

GRANT, J. et al. Postulates for revising BDI structures. **Synthese**, Springer, New York, US, v. 175, p. 39–62, 2010.

GROVE, A. Two modelings for theory change. **Journal of philosophical logic**, Springer, New York, US, v. 17, n. 2, p. 157–170, 1988.

GUERRA, P. T.; ANDRADE, A.; WASSERMANN, R. Toward the revision of ctl models through kripke modal transition systems. In: **Formal Methods: Foundations and Applications**. Berlin, DE: Springer Berlin Heidelberg, 2013. p. 115–130.

HÁJEK, A.; HARTMANN, S. Bayesian epistemology. In: DANCY, J.; SOSA, E.; STEUP, M. (Ed.). **A companion to epistemology**. Hoboken, US: John Wiley & Sons, 2010. p. 93–106.

HANSSON, S. O. In defense of base contraction. **Synthese**, Springer, New York, US, v. 91, n. 3, p. 239–245, 1992.

HERZIG, A. et al. BDI logics for BDI architectures: old problems, new perspectives. **KI-Künstliche Intelligenz**, Springer Berlin Heidelberg, Berlin, DE, p. 1–11, 2016.

HINDRIKS, K. Modules as policy-based intentions: modular agent programming in GOAL. In: **Programming Multi-Agent Systems**. Berlin, DE: Springer-Verlag, 2008. p. 156–171.

HINDRIKS, K.; VAN DER HOEK, W. Goal agents instantiate intention logic. In: **Logics in Artificial Intelligence**. New York, US: Springer, 2008. p. 232–244.

HINDRIKS, K. V. et al. Agent programming with declarative goals. In: **Intelligent Agents VII Agent Theories Architectures and Languages**. New York, US: Springer, 2001. p. 228–243.

HINDRIKS, K. V.; MEYER, J.-J. C. Toward a programming theory for rational agents. **Autonomous Agents and Multi-Agent Systems**, Springer, New York, US, v. 19, n. 1, p. 4–29, 2009.

HOWDEN, N. et al. JACK intelligent agents-summary of an agent infrastructure. In: **Proceedings of the 5th International conference on autonomous agents**. New York, US: ACM, 2001.

HUTH, M.; JAGADEESAN, R.; SCHMIDT, D. Modal transition systems: A foundation for three-valued program analysis. In: **European Symposium on Programming**. New York, US: Springer, 2001. p. 155–169.

ICARD, T.; PACUIT, E.; SHOHAM, Y. Joint revision of beliefs and intention. In: **Proceedings of the Twelfth International Conference on Principles of Knowledge Representation and Reasoning**. Menlo Park, US: AAAI Press, 2010.

JAPARIDZE, G.; DE JONGH, D. The logic of provability. **Handbook of proof theory**, North-Holland Publishing, Amsterdam, NL, v. 137, p. 475–546, 1998.

JIN, Y.; THIELSCHER, M. Iterated belief revision, revised. **Artificial Intelligence**, Elsevier, Amsterdam, NL, v. 171, n. 1, p. 1–18, 2007.

JIN, Y.; THIELSCHER, M.; ZHANG, D. Mutual belief revision: semantics and computation. In: **PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE**. Cambridge, US: MIT Press, 2007. v. 22, n. 1, p. 440.

KACI, S.; VAN DER TORRE, L. Non-monotonic reasoning with various kinds of preferences. In: **Proceedings of First Multidisciplinary Workshop on Advances in Preference Handling**. [S.l.: s.n.], 2005. p. 112–117.

KATSUNO, H.; MENDELZON, A. O. Propositional knowledge base revision and minimal change. **Artificial Intelligence**, Elsevier, Amsterdam, NL, v. 52, n. 3, p. 263–294, 1991.

KATSUNO, H.; MENDELZON, A. O. On the difference between updating a knowledge base and revising it. In: **Belief Revision**. Cambridge, UK: Cambridge University Press, 1992. p. 183–203.

KINNY, D.; GEORGE, M. Commitment and e ectiveness of situated agents. In: **Proceedings of the 12th International Joint Conference on Artificial Intelligence**. Burlington, US: Morgan Kaufmann, 1991. p. 82–88.

KONOLIGE, K.; POLLACK, M. E. A representationalist theory of intention. In: **Proceedings of the 13th International Joint Conference on Artificial Intelligence**. Burlington, US: Morgan Kaufmann, 1993. p. 390–395.

KRAUS, S.; LEHMANN, D.; MAGIDOR, M. Nonmonotonic reasoning, preferential models and cumulative logics. **Artificial intelligence**, Elsevier, v. 44, n. 1, p. 167–207, 1990.

LANG, J.; VAN DER TORRE, L.; WEYDERT, E. Hidden uncertainty in the logical representation of desires. In: **Proceedings of the 18th international joint conference on Artificial intelligence**. New York, US: Morgan Kaufmann, 2003. p. 685–690.

LEVESQUE, H. J. A logic of implicit and explicit belief. In: **Proceedings of the Fourth National Conference on Artificial Intelligence**. Palo Alto, US: AAAI Press, 1984. p. 198–202.

LIU, F. **Reasoning about preference dynamics**. New York, US: Springer, 2011.

MCCANN, H. J. Settled objectives and rational constraints. **American Philosophical Quarterly**, JSTOR, p. 25–36, 1991.

MCCARTHY, J. **Ascribing mental qualities to machines**. Stanford, EUA, 1979.

MENEGUZZI, F.; LUCK, M. Norm-based behaviour modification in BDI agents. In: **Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems**. Richland, US: International Foundation for Autonomous Agents and Multiagent Systems, 2009. p. 177–184.

MEYER, J.-J. C.; VAN DER HOEK, W.; VAN LINDER, B. A logical approach to the dynamics of commitments. **Artificial Intelligence**, Elsevier, Amsterdam, NL, v. 113, n. 1, p. 1–40, 1999.

MINTOFF, J. Rule worship and the stability of intention. **Philosophia**, Springer, New York, US, v. 31, n. 3, p. 401–426, 2004.

MOORE, G. E. Moore's paradox. In: BALDWIN, T. (Ed.). **GE Moore: Selected writings**. London, UK: Routledge, 1993. p. 207–212.

MOREIRA, A. F.; VIEIRA, R. Belief update in agentspeak-dl. In: BORDINI, R. et al. (Ed.). **Programming Multi-Agent Systems**. Dagstuhl, Germany: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2008. (Dagstuhl Seminar Proceedings, 08361).

MOREIRA, A. F. et al. Agent-oriented programming with underlying ontological reasoning. In: **Declarative Agent Languages and Technologies III**. New York, US: Springer, 2006. p. 155–170.

NAYAK, A. et al. Taking Levi identity seriously: A plea for iterated belief contraction. In: **Knowledge Science, Engineering and Management**. New York, US: Springer, 2006. p. 305–317.

PACHERIE, E. The phenomenology of action: A conceptual framework. **Cognition**, Elsevier, Amsterdam, NL, v. 107, n. 1, p. 179–217, 2008.

PLAZA, J. Logics of public communications. **Synthese**, Springer, New York, US, v. 158, n. 2, p. 165–179, 2007.

PLOTKIN, G. D. A structural approach to operational semantics. **Journal of Logic and Algebraic Programming**, v. 60, n. 61, p. 17–139, 2004.

POKAHR, A.; BRAUBACH, L.; LAMERSDORF, W. Jadex: A BDI reasoning engine. In: BORDINI, R. et al. (Ed.). **Multi-Agent Programming**. New York, US: Springer, 2005, (Multiagent Systems, Artificial Societies, and Simulated Organizations, v. 15). p. 149–174. ISBN 978-0-387-24568-3.

POLLOCK, J. L. Defeasible reasoning. **Cognitive science**, Elsevier, Amsterdam, NL, v. 11, n. 4, p. 481–518, 1987.

POLLOCK, J. L. Defeasible reasoning with variable degrees of justification. **Artificial Intelligence**, Elsevier, Amsterdam, NL, v. 133, n. 1, p. 233–282, 2001.

QUINE, W. Main trends in recent philosophy: Two dogmas of empiricism. **The Philosophical Review**, JSTOR, p. 20–43, 1951.

RAMACHANDRAN, R.; NAYAK, A. C.; ORGUN, M. A. Three approaches to iterated belief contraction. **Journal of philosophical logic**, Springer, New York, US, v. 41, n. 1, p. 115–142, 2012.

RAO, A. S. Agentspeak (l): BDI agents speak out in a logical computable language. In: **Agents Breaking Away**. New York, US: Springer, 1996. p. 42–55.

RAO, A. S.; GEORGEFF, M. P. Modeling rational agents within a BDI-architecture. In: **Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning**. San Francisco, US: Morgan Kaufmann, 1991. p. 473–484.

RAO, A. S.; GEORGEFF, M. P. BDI agents: From theory to practice. In: **Proceedings of the First International Conference on Multi-Agent Systems**. Palo Alto, US: AAAI Press, 1995. v. 95, p. 312–319.

RAO, A. S.; GEORGEFF, M. P. The semantics of intention maintenance for rational agents. In: **Proceedings of the 14th International Joint Conference on Artificial Intelligence**. San Francisco, US: Morgan Kaufmann, 1995. p. 704–710.

RAO, A. S.; GEORGEFF, M. P. Decision procedures for BDI logics. **Journal of Logic and Computation**, Oxford University Press, Oxford, UK, v. 8, n. 3, p. 293–343, 1998.

REITER, R. A logic for default reasoning. **Artificial intelligence**, Elsevier, Amsterdam, NL, v. 13, n. 1, p. 81–132, 1980.

ROTT, H. Two dogmas of belief revision. **The Journal of Philosophy**, JSTOR, v. 97, n. 9, p. 503–522, 2000.

ROTT, H. Shifting priorities: Simple representations for 27 iterated theory change operators. In: H., L.; LINDSTRÖM, S.; SLIWINSKI, R. (Ed.). **Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg**. Uppsala, SE: Uppsala Universiteit, 2006, (Uppsala Philosophical Studies, 53). p. 359–384.

ROTT, H. Shifting priorities: Simple representations for twenty-seven iterated theory change operators. **Towards Mathematical Philosophy**, Springer, New York, US, p. 269–296, 2009.

ROTT, H. 'just because': Taking belief bases seriously. In: BUSS, P. H. u. P. P. S. R. (Ed.). **Lecture Notes in Logic**. Urbana, US: Association for Symbolic Logic, 98. v. 13, p. 387–408.

ROY, O. A dynamic-epistemic hybrid logic for intentions and information changes in strategic games. **Synthese**, Springer, New York, US, v. 171, n. 2, p. 291–320, 2009.

SARDINA, S.; PADGHAM, L. A BDI agent programming language with failure handling, declarative goals, and planning. **Autonomous Agents and Multi-Agent Systems**, Springer, New York, US, v. 23, n. 1, p. 18–70, 2011.

SCHLOSSER, M. Agency. In: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Fall 2015. Stanford, US: The Metaphysics Research Lab, Stanford University, 2015. Available from Internet: <http://plato.stanford.edu/archives/fall2015/entries/agency/>. Accessed in: Nov 14 of 2016.

SCHUT, M. C. **Intention reconsideration**. Thesis (PhD) — University of Liverpool, 2002.

SEGERBERG, K. Irrevocable belief revision in dynamic doxastic logic. **Notre Dame journal of formal logic**, University of Notre Dame, v. 39, n. 3, p. 287–306, 1998.

SEGERBERG, K. Two traditions in the logic of belief: bringing them together. In: **Logic, language and reasoning**. New York, US: Springer, 1999. p. 135–147.

SEGERBERG, K. The basic dynamic doxastic logic of agm. In: **Frontiers in belief revision**. New York, US: Springer, 2001. p. 57–84.

SHOHAM, Y. Agent-oriented programming. **Artificial intelligence**, Elsevier, Amsterdam, NL, v. 60, n. 1, p. 51–92, 1993.

SHOHAM, Y. Logical theories of intention and the database perspective. **Journal of Philosophical Logic**, Springer, New York, US, v. 38, n. 6, p. 633–647, 2009.

SINGH, M. P. A critical examination of the Cohen-Levesque theory of intentions. In: **Proceedings of the 10th European conference on Artificial intelligence**. New York, US: John Wiley & Sons, 1992. p. 364–368.

SMART, J. J. C. Extreme and restricted utilitarianism. **The Philosophical Quarterly**, JSTOR, p. 344–354, 1956.

SPOHN, W. **Ordinal conditional functions: A dynamic theory of epistemic states**. New York, US: Springer, 1988.

THANGARAJAH, J.; PADGHAM, L.; HARLAND, J. Representation and reasoning for goals in BDI agents. In: **Proceedings of the twenty-fifth Australasian conference on Computer science**. Darlinghurst, AU: Australian Computer Society, 2002. v. 24, n. 1, p. 259–265.

THOMASON, R. H. Desires and defaults: A framework for planning with inferred goals. In: **Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning**. New York, US: Morgan Kaufmann, 2000. p. 702–713.

VAN BENTHEM, J. **Exploring logical dynamics**. Stanford, US: CSLI publications, 1996. 288 p. (Studies in Logic, Language, and Information, v. 156).

VAN BENTHEM, J. Modal frame correspondences and fixed-points. **Studia Logica**, Springer, New York, US, v. 83, n. 1-3, p. 133–155, 2006.

VAN BENTHEM, J. Dynamic logic for belief revision. **Journal of Applied Non-Classical Logics**, Taylor & Francis, v. 17, n. 2, p. 129–155, 2007.

VAN BENTHEM, J.; GIRARD, P.; ROY, O. Everything else being equal: A modal logic for ceteris paribus preferences. **Journal of philosophical logic**, Springer, New York, US, v. 38, n. 1, p. 83–125, 2009.

VAN BENTHEM, J.; GROSSI, D.; LIU, F. Priority structures in deontic logic. **Theoria**, John Wiley & Sons, Hoboken, US, v. 80, n. 2, p. 116–152, 2014.

VAN BENTHEM, J.; LIU, F. Dynamic logic of preference upgrade. **Journal of Applied Non-Classical Logics**, Taylor & Francis, v. 17, n. 2, p. 157–182, 2007.

VAN BENTHEM, J.; PACUIT, E.; ROY, O. Toward a theory of play: A logical perspective on games and interaction. **Games**, Molecular Diversity Preservation International, v. 2, n. 1, p. 52–86, 2011.

VAN DER HOEK, W.; JAMROGA, W.; WOOLDRIDGE, M. Towards a theory of intention revision. **Synthese**, Springer, New York, US, v. 155, n. 2, p. 265–290, 2007.

VAN DITMARSCH, H.; KOOI, B. Semantic results for ontic and epistemic change. In: **Proceedings of 7th International Conference on Logic and the Foundations of Game and Decision Theory**. New york,US: Springer, 2008. p. 87–117.

VAN LINDER, B.; VAN DER HOEK, W.; MEYER, J.-J. C. Formalising motivational attitudes of agents. In: **Intelligent Agents II Agent Theories, Architectures, and Languages**. New York, US: Springer, 1996. p. 17–32.

VAN OIJEN, J.; VANHÉE, L.; DIGNUM, F. Ciga: A middleware for intelligent agents in virtual environments. In: BEER, M. et al. (Ed.). **Agents for Educational Games and Simulations**. Berlin, DE: Springer-Verlag, 2012, (Lecture Notes in Computer Science, v. 7471). p. 22–37. Available from Internet: <http://dx.doi.org/10.1007/978-3-642-32326-3_2>. Accessed in: Nov 14 of 2016.

VAN RIEMSDIJK, M. B. et al. Dynamics of declarative goals in agent programming. In: **Declarative Agent Languages and Technologies II**. New York, US: Springer, 2005. p. 1–18.

VAN RIEMSDIJK, M. B.; DASTANI, M.; MEYER, J.-J. C. Semantics of declarative goals in agent programming. In: **Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems**. New York, US: ACM, 2005. p. 133–140.

VAN RIEMSDIJK, M. B.; DASTANI, M.; MEYER, J.-J. C. Goals in conflict: semantic foundations of goals in agent programming. **Autonomous Agents and Multi-Agent Systems**, Springer, New York, US, v. 18, n. 3, p. 471–500, 2009.

VAN ZEE, M. et al. Agm revision of beliefs about action and time. In: **Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence**. Palo Alto, US: AAAI Press, 2015.

VAN ZEE, M. et al. Consistency conditions for beliefs and intentions. In: **Proceedings of the Twelfth International Symposium on Logical Formalizations of Commonsense Reasoning**. Palo Alto, US: AAAI Press, 2015.

VELOSO, M. M.; POLLACK, M. E.; COX, M. T. Rationale-based monitoring for planning in dynamic environments. In: **Proceedings of the Fourth International Conference on Artificial Intelligence Planning Systems**. Menlo Park, US: AAAI Press, 1998. p. 171–180.

VIEIRA, R. et al. On the formal semantics of speech-act based communication in an agent-oriented programming language. **Journal of Artificial Intelligence Research**, AAAI Press, Palo Alto, US, v. 29, n. 1, p. 221–267, jun. 2007.

WILLIAMS, M.-A. On the logic of theory base change. In: **Logics in Artificial Intelligence**. New York, US: Springer, 1994. p. 86–105.

WILSON, G.; SHPALL, S. Action. In: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Summer 2012. Stanford, US: The Metaphysics Research Lab, Stanford University, 2012. Available from Internet: <http://plato.stanford.edu/archives/sum2012/entries/action/>. Accessed in: Nov 14 of 2016.

WOBCKE, W. Plans and the revision of intentions. In: **Distributed Artificial Intelligence Architecture and Modelling**. New York, US: Springer, 1996. p. 100–114.

WOBCKE, W. An operational semantics for a PRS-like agent architecture. In: **AI 2001: Advances in Artificial Intelligence**. New York, US: Springer, 2001. p. 569–580.

WOBCKE, W. Model theory for PRS-like agents: Modelling belief update and action attempts. In: **Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence**. Berlin, DE: Springer-Verlag, 2004. p. 595–604.

WOOLDRIDGE, M. Practical reasoning with procedural knowledge (a logic of BDI agents with Know-how). In: **Proceedings of the International Conference on Formal and Applied Practical Reasoning**. Berlin, DE: Springer-Verlag, 1996. (Lecture Notes in Computer Science, v. 1075), p. 202–213.

WOOLDRIDGE, M. Computationally grounded theories of agency. In: **Proceedings of the Fourth International Conference on MultiAgent Systems**. Los Alamitos, US: IEEE Computer Society, 2000. p. 13–20.

WOOLDRIDGE, M.; PARSONS, S. Intention reconsideration reconsidered. In: **Intelligent Agents V: Agents Theories, Architectures, and Languages**. New York, US: Springer, 1999. p. 63–79.

WOOLDRIDGE, M. J. **Reasoning about rational agents**. Cambridge, US: MIT press, 2000.

ZHANG, Y.; DING, Y. Ctl model update for system modifications. **Journal of Artificial Intelligence Research**, p. 113–155, 2008.

APPENDIX A — AXIOMATIZATION OF THE LOGIC OF RATIONALITY AND PRACTICALITY

Definition A.1 *Let P be a set of propositional letters and $\mathscr{A}$ a plan library, we define the language $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$ by the following grammar (where $p \in P$ and $\alpha \in \mathscr{A}$):*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid A\varphi \mid [\leq_P]\varphi \mid [<_P]\varphi \mid [\leq_D]\varphi \mid [<_D]\varphi \mid [\alpha]\varphi$$

In this logic we can enconde the mental attitudes of *knowledge*, *belief*, *overwhelming desire*, *goal*, *can achieve* and *intention*, respectively as:

$$\begin{array}{lcl}
K\psi & \equiv & A\psi \\
B(\psi|\varphi) & \equiv & A((\varphi \wedge \neg\langle<_P\rangle\varphi) \rightarrow \psi) \\
I(\psi|\varphi) & \equiv & A((\varphi \wedge \neg\langle<_D\rangle\varphi) \rightarrow \psi) \\
G(\psi|\varphi) & \equiv & E((\varphi \wedge \neg\langle<_D\rangle\varphi) \wedge \psi) \\
\diamond\varphi & \equiv & \bigvee_{\alpha\in\mathscr{A}}(pre(\alpha) \wedge [\alpha]\varphi) \\
Int(\psi|\varphi) & \equiv & I(\psi|\varphi) \wedge E(\varphi \wedge \psi) \wedge \neg B(\psi|\varphi) \wedge B(\diamond\psi|\varphi)
\end{array}$$

The dynamic logic of practical rationality $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$ (considering models with chains of maximum size n) extended with public announcements, plausibility update by radical upgrade, plausibility update by public suggestion, plausibility update by lexicographic contraction, desirability update by radical upgrade, desirability update by public suggestion and desirability update by lexicographic contraction can be axiomatized by the following axioms together with the propositional validities and the rules of *modus ponens* and necessitation.

$$
\begin{array}{lcl}
[\leq_P](\varphi \to \psi) & \to & ([\leq]\varphi \to [\leq]\psi) \\
[\leq_P]\varphi & \to & \varphi \\
[\leq_P]\varphi & \to & [\leq_P][\leq_P]\varphi
\end{array}
$$

$$
\begin{array}{lcl}
[<_P](\varphi \to \psi) & \to & ([<_P]\varphi \to [<_P]\psi) \\
[<_P]([<_P]\varphi & \to & \varphi) \to [<_P]\varphi \\
[\leq_P]\varphi & \to & [<_P]\varphi \\
[<_P]\varphi & \to & [<_P][\leq_P]\varphi \\
[<_P]\varphi & \to & [\leq_P][<_P]\varphi
\end{array}
$$

$$
\begin{array}{lcl}
[\leq_D](\varphi \to \psi) & \to & ([\leq_D]\varphi \to [\leq_D]\psi) \\
[\leq_D]\varphi & \to & \varphi \\
[\leq_D]\varphi & \to & [\leq_D][\leq_D]\varphi
\end{array}
$$

$$
\begin{array}{lcl}
[<_D](\varphi \to \psi) & \to & ([<_D]\varphi \to [<_D]\psi) \\
[<_D]([<_D]\varphi \to \varphi) & \to & [<_D]\varphi \\
[\leq_D]\varphi & \to & [<_D]\varphi \\
[<_D]\varphi & \to & [<_D][\leq_D]\varphi \\
[<_D]\varphi & \to & [\leq_D][<_D]\varphi
\end{array}
$$

$$
\begin{array}{lcl}
A(\varphi \to \psi) & \to & (A\varphi \to A\psi) \\
A\varphi & \to & \varphi \\
A\varphi & \to & AA\varphi \\
\varphi & \to & A\neg A\neg\varphi \\
A\varphi & \to & [\leq_P]\varphi \\
A\varphi & \to & [\leq_D]\varphi
\end{array}
$$

$$
\begin{array}{lcl}
[\alpha]p & \leftrightarrow & pre(\alpha) \to \top \text{ if } pos(\alpha) \vDash p \\
[\alpha]p & \leftrightarrow & pre(\alpha) \to \bot \text{ if } pos(\alpha) \vDash \neg \\
[\alpha]p & \leftrightarrow & pre(\alpha) \to p \text{ if } pos(\alpha) \nvDash p \text{ and } pos(\alpha) \nvDash \neg p \\
[\alpha](\varphi \wedge \psi) & \leftrightarrow & [\alpha]\varphi \wedge [\alpha]\psi \\
[\alpha]\neg\varphi & \leftrightarrow & pre(\alpha) \to \neg[\alpha]\varphi \\
[\alpha][\leq_P]\varphi & \leftrightarrow & pre(\alpha) \to [\leq_P][\alpha]\varphi \\
[\alpha][\leq_D]\varphi & \leftrightarrow & pre(\alpha) \to [\leq_D][\alpha]\varphi \\
[\alpha][<_P]\varphi & \leftrightarrow & pre(\alpha) \to [<_P][\alpha]\varphi \\
[\alpha][<_D]\varphi & \leftrightarrow & pre(\alpha) \to [<_D][\alpha]\varphi \\
[\alpha]A\varphi & \leftrightarrow & pre(\alpha) \to A[\alpha]\varphi
\end{array}
$$

$$
\begin{array}{lcl}
[!\varphi]p & \leftrightarrow & \varphi \to p \\
[!\varphi]\neg\psi & \leftrightarrow & \varphi \to \neg[!\varphi]\psi \\
[!\varphi]\psi \wedge \xi & \leftrightarrow & [!\varphi]\psi \wedge [!\varphi]\xi \\
[!\varphi]A\psi & \leftrightarrow & \varphi \to A([!\varphi]\psi) \\
[!\varphi][\leq_P]\psi & \leftrightarrow & \varphi \to [\leq_P][!\varphi]\psi \\
[!\varphi][<_P]\psi & \leftrightarrow & \varphi \to [<_P][!\varphi]\psi \\
[!\varphi][\leq_D]\psi & \leftrightarrow & \varphi \to [\leq_D][!\varphi]\psi \\
[!\varphi][<_D]\psi & \leftrightarrow & \varphi \to [<_D][!\varphi]\psi
\end{array}
$$

$$
\begin{array}{lcl}
[\Uparrow_P \varphi]p & \leftrightarrow & p \\
[\Uparrow_P \varphi]\neg\psi & \leftrightarrow & \neg[\Uparrow_P \varphi]\psi \\
[\Uparrow_P \varphi](\psi \wedge \xi) & \leftrightarrow & [\Uparrow_P \varphi]\psi \wedge [\Uparrow_P \varphi]\xi \\
[\Uparrow_P \varphi]A\psi & \leftrightarrow & A([\Uparrow_P \varphi]\psi) \\
[\Uparrow_P \varphi][\leq]\psi & \leftrightarrow & \neg\varphi \to (A(\varphi \to [\Uparrow_P \varphi]\psi) \wedge [\leq](\neg\varphi \to [\Uparrow_P \varphi]\psi)) \wedge \\
 & & \varphi \to [\leq](\varphi \to [\Uparrow_P \varphi]\psi) \\
[\Uparrow_P \varphi][<]\psi & \leftrightarrow & \neg\varphi \to (A(\varphi \to [\Uparrow_P \varphi]\psi) \wedge [<](\neg\varphi \to [\Uparrow_P \varphi]\psi)) \wedge \\
 & & \varphi \to [<](\varphi \to [\Uparrow_P \varphi]\psi) \\
[\Uparrow_P \varphi][\alpha]\psi & \leftrightarrow & [\alpha][\Uparrow_P \varphi]\psi \\
[\Uparrow_P \varphi][\leq_D]\psi & \leftrightarrow & [\leq_D][\Uparrow_P \varphi]\psi \\
[\Uparrow_P \varphi][<_D]\psi & \leftrightarrow & [<_D][\Uparrow_P \varphi]\psi
\end{array}
$$

$$
\begin{array}{lcl}
[\Uparrow_D \varphi]p & \leftrightarrow & p \\
[\Uparrow_D \varphi]\neg\psi & \leftrightarrow & \neg[\Uparrow_D \varphi]\psi \\
[\Uparrow_D \varphi](\psi \wedge \xi) & \leftrightarrow & [\Uparrow_D \varphi]\psi \wedge [\Uparrow_D \varphi]\xi \\
[\Uparrow_D \varphi]A\psi & \leftrightarrow & A([\Uparrow_D \varphi]\psi) \\
[\Uparrow_D \varphi][\leq]\psi & \leftrightarrow & \neg\varphi \rightarrow (A(\varphi \rightarrow [\Uparrow_D \varphi]\psi) \wedge [\leq](\neg\varphi \rightarrow [\Uparrow_D \varphi]\psi)) \wedge \\
 & & \varphi \rightarrow [\leq](\varphi \rightarrow [\Uparrow_D \varphi]\psi) \\
[\Uparrow_D \varphi][<]\psi & \leftrightarrow & \neg\varphi \rightarrow (A(\varphi \rightarrow [\Uparrow_D \varphi]\psi) \wedge [<](\neg\varphi \rightarrow [\Uparrow_D \varphi]\psi)) \wedge \\
 & & \varphi \rightarrow [<](\varphi \rightarrow [\Uparrow_D \varphi]\psi) \\
[\Uparrow_D \varphi][\alpha]\psi & \leftrightarrow & [\alpha][\Uparrow_D \varphi]\psi \\
[\Uparrow_P \varphi][\leq_D]\psi & \leftrightarrow & [\leq_P][\Uparrow_D \varphi]\psi \\
[\Uparrow_D \varphi][<_P]\psi & \leftrightarrow & [<_P][\Uparrow_D \varphi]\psi
\end{array}
$$

$$
\begin{array}{lcl}
[\#_P\varphi]p & \leftrightarrow & p \\
[\#_P\varphi]\neg\psi & \leftrightarrow & \neg[\#_P\varphi]\psi \\
[\#_P\varphi](\psi \wedge \xi) & \leftrightarrow & [\#_P\varphi]\psi \wedge [\#_P\varphi]\xi \\
[\#_P\varphi]A\psi & \leftrightarrow & A[\#_P\varphi]\psi \\
[\#_P\varphi][\leq_P]\psi & \leftrightarrow & (\varphi \rightarrow [\leq_P][\#_P\varphi]\psi) \wedge \neg\varphi \rightarrow [\leq_P](\neg\varphi \rightarrow [\#_P\varphi]\psi) \\
[\#_P\varphi][<_P]\psi & \leftrightarrow & (\varphi \rightarrow [<_P][\#_P\varphi]\psi) \wedge \neg\varphi \rightarrow [<_P](\neg\varphi \rightarrow [\#_P\varphi]\psi) \\
[\#_P\varphi][\leq_D]\psi & \leftrightarrow & [\leq_D][\#_P\varphi]\psi \\
[\#_P\varphi][<_D]\psi & \leftrightarrow & [<_D][\#_P\varphi]\psi \\
[\#_P\varphi][\alpha]\psi & \leftrightarrow & [\alpha][\#_P\varphi]\psi
\end{array}
$$

$$
\begin{array}{lcl}
[\#_D\varphi]p & \leftrightarrow & p \\
[\#_D\varphi]\neg\psi & \leftrightarrow & \neg[\#_D\varphi]\psi \\
[\#_D\varphi](\psi \wedge \xi) & \leftrightarrow & [\#_D\varphi]\psi \wedge [\#_D\varphi]\xi \\
[\#_D\varphi]A\psi & \leftrightarrow & A[\#_D\varphi]\psi \\
[\#_D\varphi][\leq_P]\psi & \leftrightarrow & [\leq_P][\#_D\varphi]\psi \\
[\#_D\varphi][<_P]\psi & \leftrightarrow & [<_P][\#_D\varphi]\psi \\
[\#_D\varphi][\leq_D]\psi & \leftrightarrow & (\varphi \rightarrow [\leq_D][\#_D\varphi]\psi) \wedge \neg\varphi \rightarrow [\leq_D](\neg\varphi \rightarrow [\#_D\varphi]\psi) \\
[\#_D\varphi][<_D]\psi & \leftrightarrow & (\varphi \rightarrow [<_D][\#_D\varphi]\psi) \wedge \neg\varphi \rightarrow [<_D](\neg\varphi \rightarrow [\#_D\varphi]\psi) \\
[\#_D\varphi][\alpha]\psi & \leftrightarrow & [\alpha][\#_D\varphi]\psi
\end{array}
$$

$\neg d_{\top}(n+1)$

$$
\begin{array}{lcl}
[\Downarrow_P \varphi]p & \leftrightarrow & p \\
[\Downarrow_P \varphi]\neg\psi & \leftrightarrow & \neg[\Downarrow_P \varphi]\psi \\
[\Downarrow_P \varphi](\xi \wedge \psi) & \leftrightarrow & [\Downarrow_P \varphi]\xi \wedge [\Downarrow_P \varphi]\psi \\
[\Downarrow_P \varphi]A\psi & \leftrightarrow & A[\Downarrow_P \varphi]\psi \\
[\Downarrow_P \varphi][\leq_D]\psi & \leftrightarrow & [\leq_D][\Downarrow_P \varphi]\psi \\
[\Downarrow_P \varphi][<_D]\psi & \leftrightarrow & [\leq_D][\Downarrow_P \varphi]\psi \\
[\Downarrow_P \varphi][\leq_P]\psi & \leftrightarrow & \varphi \rightarrow [\leq_P](\varphi \rightarrow [\Downarrow_P \varphi]\psi)\wedge \\
 & & \neg\varphi \rightarrow [\leq_P](\neg\varphi \rightarrow [\Downarrow_P \varphi]\psi)\wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i}^{n} \mu_P dg_{P\varphi}(j) \rightarrow A(\mu_P dg_{P\neg\varphi}(i) \rightarrow [\Downarrow_P \varphi]\psi)\wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i}^{n} \mu_P dg_{\neg\varphi(j)} \rightarrow A(\mu_P dg_{P\varphi}(i) \rightarrow [\Downarrow_P \varphi]\psi) \\
[\Downarrow_P \varphi][<_P]\psi & \leftrightarrow & \varphi \rightarrow [<_P](\varphi \rightarrow [\Downarrow_P \varphi]\psi)\wedge \\
 & & \neg\varphi \rightarrow [<_P](\neg\varphi \rightarrow [\Downarrow_P \varphi]\psi)\wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i+1}^{n} \mu_P dg_{P\varphi}(j) \rightarrow A(\mu_P dg_{P\neg\varphi}(i) \rightarrow [\Downarrow_P \varphi]\psi)\wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i+1}^{n} \mu_P dg_{P\neg\varphi}(j) \rightarrow A(\mu_P dg_{P\varphi}(i) \rightarrow [\Downarrow_P \varphi]\psi)
\end{array}
$$

$$
\begin{array}{lcl}
[\Downarrow_D \varphi]p & \leftrightarrow & p \\
[\Downarrow_D \varphi]\neg\psi & \leftrightarrow & \neg[\Downarrow_D \varphi]\psi \\
[\Downarrow_D \varphi](\xi \wedge \psi) & \leftrightarrow & [\Downarrow_D \varphi]\xi \wedge [\Downarrow_D \varphi]\psi \\
[\Downarrow_D \varphi]A\psi & \leftrightarrow & A[\Downarrow_D \varphi]\psi \\
[\Downarrow_D \varphi][\leq_P]\psi & \leftrightarrow & [\leq_P][\Downarrow_D \varphi]\psi \\
[\Downarrow_D \varphi][<_P]\psi & \leftrightarrow & [<_P][\Downarrow_D \varphi]\psi \\
[\Downarrow_D \varphi][\leq_D]\psi & \leftrightarrow & \varphi \rightarrow [\leq_D](\varphi \rightarrow [\Downarrow_D \varphi]\psi)\wedge \\
 & & \neg\varphi \rightarrow [\leq_D](\neg\varphi \rightarrow [\Downarrow_D \varphi]\psi)\wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i}^{n} \mu_D dg_{D\varphi}(j) \rightarrow A(\mu_D dg_{D\neg\varphi}(i) \rightarrow [\Downarrow_D \varphi]\psi)\wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i}^{n} \mu_D dg_{D\neg\varphi(j)} \rightarrow A(\mu_D dg_{D\varphi}(i) \rightarrow [\Downarrow_D \varphi]\psi) \\
[\Downarrow_D \varphi][<_D]\psi & \leftrightarrow & \varphi \rightarrow [<_D](\varphi \rightarrow [\Downarrow_D \varphi]\psi)\wedge \\
 & & \neg\varphi \rightarrow [<_D](\neg\varphi \rightarrow [\Downarrow_D \varphi]\psi)\wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i+1}^{n} \mu_D dg_{D\varphi}(j) \rightarrow A(\mu_D dg_{D\neg\varphi}(i) \rightarrow [\Downarrow_D \varphi]\psi)\wedge \\
 & & \bigwedge_{i=1}^{n}\bigwedge_{j=i+1}^{n} \mu_D dg_{D\neg\varphi}(j) \rightarrow A(\mu_D dg_{D\varphi}(i) \rightarrow [\Downarrow_D \varphi]\psi)
\end{array}
$$

APPENDIX B — RESUMO ESTENDIDO

Dada a importância de agentes inteligentes e sistemas multiagentes na Ciência da Computação e na Inteligência Artificial, a programação orientada a agentes (AOP, do inglês *Agent-oriented programming*) emergiu como um novo paradigma para a criação de sistemas computacionais complexos. Assim, nas últimas décadas, houve um florescimento da literatura em programação orientada a agentes e, com isso, surgiram diversas linguages de programação seguindo tal paradigma, como AgentSpeak (RAO, 1996; BORDINI; HUBNER; WOOLDRIDGE, 2007), Jadex (POKAHR; BRAUBACH; LAMERSDORF, 2005), 3APL/2APL (DASTANI; VAN RIEMSDIJK; MEYER, 2005; DASTANI, 2008), GOAL (HINDRIKS et al., 2001), entre outras.

Programação orientada a agentes é um paradigma de programação proposto por Shoham (1993) no qual os elementos mínimos de um programa são agentes. Shoham (1993) defende que agentes autônomos e sistemas multiagentes configuram-se como uma forma diferente de se organizar uma solução para um problema computacional, de forma que a construção de um sistema multiagente para a solução de um problema pode ser entendida como um paradgima de programação.

Para entender tal paradigma, é necessário entender o conceito de agente. Agente, nesse contexto, é uma entidade computacional descrita por certos atributos - chamados de atitudes mentais - que descrevem o seu estado interno e sua relação com o ambiente externo. Atribuir a interpretação de atitudes mentais a tais atributos é válida, defende Shoham (1993), uma vez que esses atributos se comportem de forma semelhante as atitudes mentais usadas para descrever o comportamento humano e desde que sejam pragmaticamente justificáveis, i.e. úteis à solução do problema.

Entender, portanto, o significado de termos como 'crença', 'desejo', 'intenção', etc., assim como suas propriedades fundamentais, é de fundamental importância para estabelecer linguagens de programação orientadas a agentes. Nesse trabalho, vamos nos preocupar com um tipo específico de linguagens de programação orientadas a agentes, as chamadas linguagens BDI. Linguagens BDI são baseadas na teoria BDI da Filosofia da Ação em que o estado mental de um agente (e suas ações) é descrito por suas crenças, desejos e intenções.

Enquanto a construção de sistemas baseados em agentes e linguagens de programação foram tópicos bastante discutidos na literatura, a conexão entre tais sistemas e linguagens com o trabalho teórico proveniente da Inteligência Artificial e da Filosofia da Ação ainda não está bem estabelecida. Essa distância entre a teoria e a prática da construção de sistemas é bem reconhecida na literatura relevante e comumente chamada de "*gap* semântico" (*gap* em inglês

significa lacuna ou abertura e representa a distância entre os modelos teóricos e sua implementação em linguagens e sistemas).

Muitos trabalhos tentaram atacar o problema do *gap* semântico para linguagens de programação específicas, como para as linguagens AgentSpeak (BORDINI; MOREIRA, 2004), GOAL (HINDRIKS; VAN DER HOEK, 2008), etc. De fato, Rao (1996, p. 44) afirma que "O cálice sagrado da pesquisa em agentes BDI é mostrar uma correspondência 1-a-1 com uma linguagem razoavelmente útil e expressiva" (tradução nossa)[1]

Uma limitação crucial, em nossa opinião, das tentativas passadas de estabeler uma conexão entre linguagens de programação orientadas a agentes e lógicas BDI é que elas se baseiam em estabelecer a interpretação de um programa somente no nível estático. De outra forma, dado um estado de um programa, tais trabalhos tentam estabelecer uma interpretação declarativa, i.e. baseada em lógica, do estado do programa respresentando assim o estado mental do agente. Não é claro, entretanto, como a execução do programa pode ser entendida enquanto mudanças no estado mental do agente.

A razão para isso, nós acreditamos, está nos formalismos utilizados para especificar agentes BDI. De fato, as lógicas BDI propostas são, em sua maioria, estáticas ou incapazes de representar ações mentais.

O ato de revisão uma crença, adotar um objetivo ou mudar de opinião são exemplos de ações mentais, i.e. ações que são executadas internarmente ao agente e afetando somente seu estado mental, sendo portanto não observáveis. Tais ações são, em nossa opinião, intrinsecamente diferentes de ações ônticas que consistem de comportamento observável e que possivelmente afeta o ambiente externo ao agente.

Essa diferença é comumente reconhecida no estudo da semântica de linguagens de programação orientadas a agentes (BORDINI; HUBNER; WOOLDRIDGE, 2007; D'INVERNO et al., 1998; MENEGUZZI; LUCK, 2009), entretanto os formalismos disponíveis para se especificar raciocínio BDI, em nosso conhecimento, não provem recursos expressivos para codificar tal diferença. Nós acreditamos que, para atacar o *gap* semântico, precisamos de um ferramental semântico que permita a especificação de ações mentais, assim como ações ônticas.

Lógicas Dinâmicas Epistêmicas (DEL, do inglês *Dynamic Epistemic Logic*) são uma família de lógicas modais dinâmicas largamente utilizadas para estudar os fenômenos de mudança do estado mental de agentes. Os trabalhos em DEL foram fortemente influenciados pela escola holandesa de lógica, com maior proponente Johna Van Benthem, e seu "desvio dinâmico" em lógica (*dynamic turn* em inglês) que propõe a utilização de lógicas dinâmicas

[1]No original, em inglês: *"[t]he holy grail of BDI agent research is to show such a one-to-one correspondence with a reasonably useful and expressive language."*

para compreender ações de mudanças mentais (VAN BENTHEM, 1996).

O formalismo das DEL deriva de diversas vertentes do estudo de mudança epistêmica, como o trabalho em teoria da Revisão de Crenças AGM (ALCHOURRÓN; GÄRDENFORS; MAKINSON, 1985), e Epistemologia Bayesiana (HÁJEK; HARTMANN, 2010). Tais lógicas adotam a abordagem, primeiro proposta por Segerberg (1999), de representar mudanças epistêmicas dentro da mesma linguagem utilizada para representar as noções de crença e conhecimento, diferente da abordagem extra-semântica do Revisão de Crenças *a la* AGM.

No contexto das DEL, uma lógica nos parece particulamente interessante para o estudo de programação orientada a agentes: a Lógica Dinâmica de Preferências (DPL, do inglês *Dynamic Preference Logic*) de Girard (2008). DPL, também conhecida como lógica dinâmica de ordem, é uma lógica dinâmica para o estudo de preferências que possui grande expressibilidade para codificar diversas atiutudes mentais. De fato, tal lógica foi empregada para o estudo de obrigações (VAN BENTHEM; GROSSI; LIU, 2014), crenças (GIRARD; ROTT, 2014), preferências (GIRARD, 2008), etc. Tal lógica possui fortes ligações com raciocínio não-monotônico e com lógicas já propostas para o estudo de atitudes mentais na área de Teoria da Decisão (BOUTILIER, 1994b)

Nós acreditamos que DPL constitui um candidato ideal para ser utilizado como ferramental semântico para se estudar atitudes mentais da teoria BDI por permitir grande flexibilidade para representação de tais atitudes, assim como por permitir a fácil representação de ações mentais como revisão de crenças, adoção de desejos, etc. Mais ainda, pelo trabalho de Liu (2011), sabemos que existem representações sintáticas dos modelos de tal lógica que podem ser utilizados para raciocinar sobre atitudes mentais, sendo assim candidatos naturais para serem utilizados como estruturas de dados para uma implementação semanticamente fundamentada de uma linguagem de programação orientada a agentes.

Assim, nesse trabalho nós avançamos no problema de reduzir o *gap* semântico entre linguagens de programação orientadas a agentes e formalismos lógicos para especificar agentes BDI. Nós exploramos não somente como estabelecer as conexões entre as estruturas estáticas, i.e. estado de um programa e um modelo da lógica, mas também como as ações de raciocínio pelas quais se especifica a semântica formal de uma linguagem de programação orientada a agentes podem ser entendidas dentro da lógica como operadores dinâmicos que representam ações mentais do agente. Com essa conexão, nós provemos também um conjunto de operações que podem ser utilizadas para se implementar uma linguagem de programação orientada a agentes e que preservam a conexão entre os programas dessa linguagem e os modelos que representam o estado mental de um agente.

Finalmente, com essas conexões, nós desenvolvemos um arcabouço para estudar a dinâmica de atitudes mentais, tais como crenças, desejos e inteções, e como reproduzir essas propriedades na semântica de linguagens de programação.

B.1 Intenções

No estudo de ações e agência, um conceito central é o de intenção. Existem na literatura filosófica diversas propostas para o significado do termo 'intenção.' Particularmente interessante para nosso estudo é a proposta de Bratman (1999) que sugere que intenções são atitudes mentais *sui generis* que controlam o comportamento de um determinado agente.

A teoria da ação de Bratman postula que a ação de um agente racional pode ser descrita por meio de três atitudes mentais principais - crenças, desejos e intenções. Tal teoria ficou conhecida como teoria BDI (do inglês *Belief-Desire-Intention*, significando Crença-Desejo-Intenção) e teve profunda influência na Inteligência Artificial.

Do ponto de vista da Inteligência Artificial, o primeiro estudo de que temos conhecimento sobre agentes que segue o paradigma BDI é o trabalho de Cohen and Levesque (1990). Os autores propõem uma codificação da conceito de intenção de Bratman em uma lógica multimodal. Para tanto, os autores elencam sete requisitos que consideram centrais no trabalho de Bratman (1999) para entender o conceito de intenções. Tais requisitos ficaram conhecido como a *desiderata* de Cohen e Levesque.

O trabalho de Cohen and Levesque (1990) estimulou muita pesquisa em Inteligência Artificial sobre agentes, culminando no paradigma de programação orientado a agentes, porposto por Shoham (1993), e na proposta de diversas linguagens de programação seguindo tal paradigma, e.g. AgentSpeak (RAO, 1996), Jadex (POKAHR; BRAUBACH; LAMERSDORF, 2005), 3APL (DASTANI; VAN RIEMSDIJK; MEYER, 2005), entre outras.

Enquanto tais trabalhos focaram no conceito de intenções e sua relação com outras atitudes mentais, poucos se concentraram no problema de quando e como um agente muda sua intenções. Bratman (1999) lista a *estabilidade* de intenções como uma propriedade fundamental dessas, uma vez que uma intenção tem o papel de reduzir as opções de ação que um agente considera em qualquer momento. O autor, entretanto, desenvolve muito pouco quais os critérios gerais para se revisar uma intenção.

A resposta de Bratman (1992) sobre como e quando reconsiderar intenções consiste em estabelecer que cada agente possui uma política, ou regra geral, que estabelece quando ele deve reconsiderar uma dada intenção. Tal resposta assemelha-se bastante, entretanto, com

a proposta de utilitarismo baseado em regras e, portanto, está sujeita a críticas semelhantes (SMART, 1956).

Mintoff (2004), estudando uma abordagem preservacionista de intenções influenciada pelo utilitarismo baseado em regras de Bratman (1992), propõe critérios positivos para a reconsideração de uma intenção, justificando-os através de um enfoque utilitarista. Tais critérios assemelham-se bastante ao que Castelfranchi and Paglieri (2007), em sua teoria construtiva de intenções, requer para a reconsideração de uma dada intenção. Consideramos que tal resposta é uma proposta clara e bem motivada para reconsideração de intenções e utilizamos tais critérios para avaliar o comportamento dinâmico de intenções nos formalismos que propomos nesse trabalho.

B.2 Uma lógica dinâmica para crenças e desejos

Nesse trabalho, utilizaremos o formalismo da Lógica Dinâmica de Preferência (DPL, do inglês *Dynamic Preference Logic*) proposta por Girard (2008). Lógica Dinâmica de Preferência é uma lógica modal dinâmica pertencente á familia das Lógicas Dinâmicas Epistêmicas. Tal lógica foi bastante utilizada para estudar codificações de diversas atitudes mentais e seu comportamento dinâmico, como crenças (BALTAG; MOSS; SOLECKI, 1998), desejos (BOUTILIER, 1994b), obrigações (VAN BENTHEM; GROSSI; LIU, 2014), etc.

Nós propomos uma lógica $\mathscr{L}_{\leq_P,\leq_D}(P)$ para representar o estado doxástico e conativo de um determinado agente, similar à lógica proposta por Boutilier (1994b). Tal lógica consiste em se acrescentar dois modais de ordem, tal qual estudado por Girard (2008) em DPL, à lógica proposicional clássica. Assim, obtemos a seguinte linguagem.

Definition B.1 *Nós definimos a linguagem $\mathscr{L}_{\leq_P,\leq_D}(P)$ pela seguinte gramática (em que $p \in P$ um conjunto finito de simbolos proposicionais):*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid A\varphi \mid [\leq_P]\varphi \mid [<_P]\varphi \mid [\leq_D]\varphi \mid [<_D]\varphi$$

Como usual na literatura, definimos as fórmulas $E\varphi \equiv \neg A\neg\varphi$ e $\langle\leq\rangle\varphi \equiv \neg[\leq]\neg\varphi$. A fórmula $[\leq_D]\varphi$ ($[\leq_P]\varphi$) significa que em todos os mundos igualmente ou mais desejáveis (plausíveis) que o atual, φ vale e $[<_D]\varphi$ ($[<_P]\varphi$) significa que em todos os mundo possíveis estritamente mais desejáveis (plausíveis) que o atual, φ vale.

Tais fórmulas são interpretadas atrabés de um modelo de Kripke contendo duas relações de acessibildiade - uma para plausibilidade e uma para desejabilidade. Nós chamamos esse tipo

de modelo de um modelo de agente

Definition B.2 *Um modelo de agente é uma tupla $M = \langle W, \leq_P, \leq_D, v\rangle$ onde W é um conjunto de mundos possíveis e ambos $\leq_D$ e $\leq_P$ são pré-ordens sobre W com partes estritas bem-fundadas.*

A interpretação das fórmulas da lógica podem então ser construídas de forma usual.

$$
\begin{aligned}
M, w \vDash [\leq_P]\varphi \quad & \text{sse } \forall w' \in W : w' \leq_P w \Rightarrow M, w' \vDash \varphi \\
M, w \vDash [<_P]\varphi \quad & \text{sse } \forall w' \in W : w' <_P w \Rightarrow M, w' \vDash \varphi \\
M, w \vDash [\leq_D]\varphi \quad & \text{sse } \forall w' \in W : w' \leq_D w \Rightarrow M, w' \vDash \varphi \\
M, w \vDash [<_D]\varphi \quad & \text{sse } \forall w' \in W : w' <_D w \Rightarrow M, w' \vDash \varphi
\end{aligned}
$$

Dentro dessa lógica, codificamos as noções de 'sabe-se que φ', 'dado φ, crê-se que ψ', 'dado φ, necessariamente deseja-se que ψ' e 'dado φ, tem-se um desejo de que ψ' através das fórmulas condicionais $K(\varphi)$, $B(\psi|\varphi)$, $I(\psi|\varphi)$ e $G(\psi|\varphi)$, respectivamente.

$$
\begin{aligned}
K(\varphi) &\equiv A\varphi \\
B(\psi|\varphi) &\equiv A((\varphi \wedge \neg\langle <_P\rangle\varphi) \rightarrow \psi) \\
I(\psi|\varphi) &\equiv A((\varphi \wedge \neg\langle <_D\rangle\varphi) \rightarrow \psi) \\
G(\psi|\varphi) &\equiv E((\varphi \wedge \neg\langle <_D\rangle\varphi) \wedge \psi)
\end{aligned}
$$

Baseando-nos no trabalho de Bratman (1999) e sua codificação lógica na teoria de Cohen and Levesque (1990), entretanto, sabemos o conceito de intenção está intimamente ligado ao de ação. Estedemos então nossa lógica $\mathscr{L}_{\leq_P,\leq_D}(P)$ para incluir ações ônticas, i.e. ações que afetam o mundo exterior ao agente, obtendo então a lógica $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$, onde $\mathscr{A}$ é uma biblioteca de ações, baseada no trabalho sobre ações ônticas em Lógica Dinâmica Epistêmica de Van Ditmarsch and Kooi (2008).

Tal lógica acresce a linguagem de $\mathscr{L}_{\leq_P,\leq_D}(P)$ com formulas do tipo $[\alpha]\varphi$ em que α é uma ação ôntica de $\mathscr{A}$. Tais formulas possuem o significado de 'após a realização da ação α, φ vale'.

Com tal lógica expandida, nós representamos as noções de 'é possível alcançar um estado em que φ vale' e'dado φ, intenciona-se que ψ valha' através das fórmulas $\diamond\varphi$ e $Int(\psi|\varphi)$, respectivamente.

$$
\begin{aligned}
\diamond\varphi &\equiv \bigvee_{\alpha\in\mathscr{A}}(pre(\alpha) \wedge [\alpha]\varphi) \\
Int(\psi|\varphi) &\equiv I(\psi|\varphi) \wedge E(\varphi \wedge \psi) \wedge \neg B(\psi|\varphi) \wedge B(\diamond\psi|\varphi)
\end{aligned}
$$

Por fim, nós incluímos nessa lógica operações dinâmicas tais quais as estudadas por Girard (2008) e na área de Revisão de Crenças (RAMACHANDRAN; NAYAK; ORGUN, 2012).

Tais operações representam operações mentais comuns ao raciocínio de um agente como *atualizar as crenças por revisão radical*, ou *atualizar os desejos por sugestão pública*, etc.

B.3 AAP: uma linguagem de programação abstrata para agentes

Nós porpomos, então, uma linguagem de programação abstrata orientada a agentes chamada AAP. Tal linguagem possui os componentes principais presentes nas principais linguagens de programação disponíveis na literatura.

Um programa em AAP pode ser identificado com uma estrutura $ag = \langle K, B, G, I \rangle$ definida sobre uma base de planos Π.

Definition B.3 *Seja P um conjunto finito de variáveis proposicionais e* Π *uma base de planos. Nós chamamos de programa de agente (agente ou estado de um agente) AAP sobre* Π*, a tupla* $ag = \langle K, B, G, I \rangle$ *em que:*

- $K \subset \mathscr{L}_O(P)$*, é um conjunto finito e consistente de formulas proposicionais, chamado base de conhecimento;*
- $B \subset \mathscr{L}_O(P) \times \mathbb{N}^*$ *é um conjunto finito de pares* $\langle \varphi, i \rangle$*, chamado base de crenças estratificada, em que* φ *é uma fórmula proposicional e i um número natural, chamado de plausibilidade ou rank de* φ *em B.*
- $G \subset \mathscr{L}_O(P) \times \mathbb{N}^*$ *é um conjunto finito de pares* $\langle \varphi, i \rangle$*, chamado base de objetivos estratificada, em que* φ *é uma fórmula proposicional e i um número natural, chamado de desejabilidade ou rank de* φ *em G.*
- $I \subset \mathscr{H}_\Pi$ *é um conjunto de planos de* Π*, chamado base de intenções procedurais.*

Utilizando a estrutura de um programa AAP, nós provemos definições do que significa um agente AAP $ag = \langle K, B, G, I \rangle$ ter um determinado conhecimento, crença, objetivo ou intenção.

Adicionalmente, nós mostramos que um agente AAP $ag = \langle K, B, G, I \rangle$ pode ser convertido em uma estrutura $\mathscr{G}_{ag} = \langle \mathscr{G}_P, \mathscr{G}_D \rangle$ chamada estrutura de agente, em que $\mathscr{G}_P$ e $\mathscr{G}_D$ são grafos de prioridades, tal qual definido por Liu (2011). Usando um resultado de Liu (2011) sobre a conexão de grafos de prioridade e modelos de DPL, nós podemos transformar a estrutura de agente $\mathscr{G}_{ag}$ em um modelo de agentes em que as noções de conhecimento, crença, objetivo e intenção coincidem com a noção dada em AAP. De outra forma, o agente *ag* sabe (crê/objetiva/intenciona) uma fórmula φ sse a fórmula $K\varphi$ ($B\varphi$/$G\varphi$/$Int\varphi$) é válida no modelo obtido a partir de $\mathscr{G}_{ag}$.

Além dessa conexão estática, nós mostramos que as regras da semântica operacional da linguagem AAP podem ser entendidas como a aplicação das operações dinâmicas definidas na lógica $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$ sobre o modelo representando o estado mental do agente.

Assim, nós mostramos uma conexão forte entre a semântica da linguagem de programação abstrata AAP com modelos formais de agentes numa determinada lógica. Com isso, podemos utilizar a lógica $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$ para raciocinar sobre a execução de programas AAP, assim como utilizar programas AAP para executar tarefas de raciocínio sobre a lógica $\mathscr{L}_{\leq_P,\leq_D}(P,\mathscr{A})$.