MONOGRAFIES DE L'INSTITUT D´INVESTIGACIÓ EN INTEL·LIGÈNCIA ARTIFICIAL

# Coherence-Based Computational Agency

Sindhu Joseph

Consell Superior d´Investigacions Científiques

MONOGRAFIA DE L'INSTITUT D'INVESTIGACIÓ
EN INTEL·LIGÈNCIA ARTIFICIAL
Number 45



# Coherence-Based Computational Agency

Sindhu Joseph

Foreword by Prof. Carles Sierra and Dr. Marco Schorlemmer

2011 Consell Superior d'Investigacions Científiques
Institut d'Investigació en Intel·ligència Artificial
Bellaterra, Catalonia, Spain.

Series Editor
Institut d'Investigació en Intel·ligència Artificial
Consell Superior d'Investigacions Científiques
Foreword by
Prof. Carles Sierra and Dr. Marco Schorlemmer
Institut d'Investigació en Intel·ligència Artificial
Consell Superior d'Investigacions Científiques
Volume Author
Sindhu Joseph
Institut d'Investigació en Intel·ligència Artificial
Consell Superior d'Investigacions Científiques





ISBN: 978-84-00-09354-9
ISBN (online): 978-84-00-09355-6
NIPO: 472-11-146-4
NIPO (online): 472-11-145-9
D.L.: B.33337-2011


**Ordering Information:** Text orders should be addressed to the Library of the IIIA, Institut d'Investigació en Intel·ligència Artificial, Campus de la Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.

*To Ammachi and Chachan*

# Contents

**Foreword** **xiii**

**I Setting the Stage** **1**

**1 Introduction** **3**
1.1 Motivations . . . 3
1.1.1 Autonomous Agents . . . 5
1.1.2 Autonomous Normative Agents and Normative MAS . . . 8
1.2 Contributions . . . 9
1.2.1 Autonomous Agents . . . 10
1.2.2 Autonomous Normative Agents and Normative MAS . . . 11
1.2.3 Other Contributions . . . 12
1.3 Organisation of the Thesis . . . 12

**2 Background** **15**
2.1 Motivational Theories of Agency . . . 15
2.1.1 Intentional Stance . . . 16
2.1.2 Utility Maximisation . . . 17
2.1.3 Reinforcement Learning . . . 18
2.1.4 Theory of Coherence . . . 19
2.1.5 Comparison with Other Decision Theories . . . 21
2.1.6 In the Context of Philosophical Theories . . . 21
2.2 Computational Formalisms . . . 22
2.2.1 The Soar Architecture . . . 22
2.2.2 BDI family of Architectures . . . 23
2.2.3 A BDI Architecture with Coherence . . . 26
2.3 Applications in normative MAS . . . 27
2.3.1 Normative Reasoning . . . 28
2.3.2 Multiagent Norm Deliberation . . . 29
2.4 Other Applications . . . 31
2.4.1 Coherence and Linguistic Analysis . . . 31
2.4.2 Coherence and Legal Reasoning . . . 32
2.5 Concluding Remarks . . . 32

**II Framework and Architecture** **35**

**3 Coherence Framework and Deductive Coherence** **37**
3.1 Generic Coherence Framework . . . . . 37
3.1.1 Coherence Graphs . . . . . 38
3.1.2 Calculating Coherence . . . . . 39
3.2 Deductive Coherence . . . . . 41
3.2.1 Deductive Coherence Function . . . . . 42
3.3 Discussion . . . . . 47

**4 Formalising Coherence: A Proof-Theoretical Approach** **49**
4.1 Properties of Deductive Coherence Based on MDRs . . . . . 49
4.1.1 Combining Conjunction . . . . . 50
4.1.2 Internal Conjunction . . . . . 50
4.1.3 Combining Disjunction . . . . . 51
4.1.4 Internal Disjunction . . . . . 51
4.1.5 Combining Implication . . . . . 51
4.1.6 Internal Implication . . . . . 52
4.1.7 Internal Negation . . . . . 52
4.2 Concluding Remarks . . . . . 52

**5 Coherence-driven Agents** **55**
5.1 Organisation of the Chapter . . . . . 55
5.2 Adaptation of MCS . . . . . 56
5.2.1 Cognitive Contexts . . . . . 57
5.2.2 Reasoning Across Contexts . . . . . 59
5.3 Coherence-Driven Agents . . . . . 64
5.4 Discussion . . . . . 66

**6 Coherence-driven Agents—Experimental Evaluation** **67**
6.1 Experimental Evaluation . . . . . 67
6.1.1 Experimental Set-up . . . . . 68
6.1.2 Variables . . . . . 69
6.1.3 Hypothesis . . . . . 70
6.2 Design of Players . . . . . 71
6.3 Simulations . . . . . 74
6.4 Discussion . . . . . 76

**III Extensions and Applications** **79**

**7 Autonomous Normative Agents** **81**
7.1 Example — Norm Deliberation . . . . . 82
7.2 Norms . . . . . 84
7.3 Extending the Coherence Framework . . . . . 85
7.4 Reasoning about Norms . . . . . 88

7.4.1 Norm Generation . . . . . 88
7.4.2 Generating a Norm Proposal . . . . . 91
7.4.3 Evaluating a Norm Proposal . . . . . 91
7.5 Discussion . . . . . 94

**8 Multiagent Norm Deliberation** **95**
8.1 Dialogue System . . . . . 96
8.1.1 Assumptions . . . . . 97
8.1.2 Communication and Topic Languages . . . . . 98
8.1.3 Joint Coherence Graph $\mathcal{J}$ . . . . . 99
8.1.4 Protocol . . . . . 103
8.2 Comparison with Other Argumentation Systems . . . . . 108
8.3 Discussion . . . . . 110

**9 Coherence: When is it Right?** **111**
9.1 Background . . . . . 112
9.1.1 Rationality . . . . . 112
9.1.2 Preference Relation . . . . . 113
9.2 Utility Coherence Graphs . . . . . 113
9.3 Dynamism in Preference Ordering . . . . . 117
9.4 Discussion . . . . . 119

**IV Conclusion and Future Directions** **121**

**10 Concluding Remarks** **123**
10.1 Autonomous Agent Reasoning . . . . . 123
10.2 Normative MAS . . . . . 125
10.3 Theory of Coherence . . . . . 126
10.4 Future directions . . . . . 127
10.4.1 Formalisation of Coherence . . . . . 127
10.4.2 Coherence and Autonomous Agents . . . . . 128
10.4.3 Coherence and Normative MAS . . . . . 129
10.4.4 Applications . . . . . 130
10.4.5 Terminology . . . . . 131
10.4.6 Pre-treaty situation (1891) . . . . . 131
10.4.7 The Incoherence Buildup (1991) . . . . . 133

# List of Figures


3.1 An example of a coherence graph . . . . . 38
3.2 A partition of a coherence graph . . . . . 40
3.3 Coherence-maximising partition of a coherence graph . . . . . 41

5.1 Architecture of a coherence-driven agent . . . . . 65

6.1 Grid Environment . . . . . 68
6.2 Coherence graph of agent $e_s = 0.8, d_s = 0.45, p_s = 1$ . . . . . 73
6.3 Simulation statistics . . . . . 74
6.4 Result of Quantile-Quantile test showing the variables in a normal distribution . . . . . 75
6.5 Tukey test to compare performance of player types *near-opt* (0), *coherence* (1), and *human* (2) . . . . . 76
6.6 Tukey test to conduct a pairwise comparison of the performance under different density change frequencies $(0, 20, 50$ and $80)$. . . . 77

7.1 The support set of node $(B\alpha, 1)$ of agent $a$ . . . . . 87
7.2 The conflict set of node $(Bc, 0.8)$ of agent $a$ . . . . . 87
7.3 The coherence graph of agent $a$ with updated norm information . 90
7.4 Norm proposal of agent $a$ . . . . . 92
7.5 Initial Coherence graph of agent $b$ with the generated norms . . . 93
7.6 Evaluation of norm proposal of agent $a$ by agent $b$ . . . . . 93

8.1 Joint coherence graph $\mathcal{J}_d$ with $d = m_1$ . . . . . 101
8.2 Joint coherence graph $\mathcal{J}_d$ with $d = m_1, m_2$ (a coherence maximising partition) . . . . . 102
8.3 Joint coherence graph $\mathcal{J}_d$ with $d = m_1, m_2$ (another coherence maximising partition) . . . . . 102
8.4 $\mathcal{J}_d$ when $d = m_1, m_2, m_3$ . . . . . 105
8.5 Internal coherence graph of agent $b$ when $d = m_1, m_2, m_3$ (agent $a$'s proposal are the shaded nodes) . . . . . 106
8.6 Joint graph when $d = m_1, m_2, m_3, m_4$ and the preferred partitions of $a$ an $b$ . . . . . 107

9.1 An example of a utility coherence graph given $o_1 \succ o_2 \succ o_3 \succ 0_4$ 116

9.2 The utility coherence graph of *proposer* in ultimatum game . . . 117
9.3 The joint coherence graph of *proposer* with $(B5, 1)$ added. . . . 119

10.1 Initial coherence graph $(g_1)$ of $s$ as in 1891 including the bridge rule deductions (shadowed) with $\kappa(g_1) = 0.32$ . . . . . 132
10.2 Coherence graph $(g_2)$, with norm accepted $\kappa(g_1) = 0.225$ . . . . . 134
10.3 Subgraph of the coherence graph $(g_3)$ . . . . . 135

# Foreword

As scientists, it is always encouraging to have the opportunity to participate in scientific inquiry that aspires to go beyond one's own narrow field of expertise, and that challenges one's conventional way of thinking, thus opening up new and exciting lines of scientific study. This is, precisely, what happend while supervising the doctoral research reported in this book, because its author, Sindhu Joseph, chose to venture into several varying areas of knowledge.

If you value interdisciplinary research based on a combination of techniques that draws from different research disciplines, you are going to enjoy reading this book. Sindhu Joseph, namely, had the insight of taking the general theory of coherence as proposed by Canadian philosopher Paul Thagard in order to extend present-day agent architectures based on the Belief-Desire-Intention model for rational reasoning and decision-making, hence combining theories from philosophy and mathematical logic with those from cognitive science and computer science. But Sindhu Joseph's book not only provides a formal ground for deductive coherence and its relation to classical logical consequence, it also describes a concrete computational framework in which these more abstract and theoretical ideas are put into practice: she has explored how to apply coherence-based reasoning to argumentation theory, and even illustrates how coherence-maximisation can be a valid alternative model —and probably a more cognitively grounded one— for rationality, in contrast to utility-maximisation as found in neoclassical economics.

All in all, this book makes an interesting read for all those who look for more cognitively-inspired computational paradigms of intelligent behaviour together with concrete implementations in application domains such as multi-agent systems and computational argumentation theory. If this is your case, you will definitely find inspiration in these pages for your own research agenda.

Bellaterra, October 2011

Carles Sierra and Marco Schorlemmer
Artificial Intelligence Research Institute, IIIA-CSIC

# Abstract

In this book we address the problem of introducing flexibility and adaptability in autonomous agent design in the context of agents situated in regulated environments. We argue that current cognitive architectures such as those based on BDI theory fall short in performing autonomous reasoning in agents. One of the important reasons is the lack of clear motivations for choosing a goal or an action to pursue. Instead of the intention-driven philosophy in a BDI architecture, we need a formalism which would dynamically select goals and intentions considering constraints among cognitive elements. Hence, our central thesis in this book is that agent architectures need to incorporate a motivational criterion which can be computed in terms of their cognitive elements while preserving those formal properties that make BDI-based architectures attractive. This book proposes the cognitive theory of coherence as one such motivational criterion for agents to reason and take autonomous decisions.

The cognitive coherence theory we use is the one proposed by Paul Thagard. The term *coherence* is defined as the quality or the state of cohering, especially a logical, orderly, and aesthetically consistent relationship of parts. A coherent set is interdependent such that every *piece of information* in it contributes to the coherence. We take Thagard's proposal of coherence as that of maximising satisfaction of constraints between pieces of information and put to use in the design of autonomous agents. This book advances the state of the art by proposing a computational formalisation of coherence that includes a mechanism to compute coherence values between pairs of pieces of information by formalising *deductive coherence*.

A central contribution in this book is the proposal of a coherence-based agent architecture which extends a BDI architecture with the notion of coherence. This architecture while preserving those formal properties of BDI-based architectures, incorporates coherence as the central motivational drive and reasons under uncertainty. Based on experimental evaluation, we prove the feasibility of coherence-driven agents and show that their performance match that of humans. We further extend coherence-based architecture for normative agents to reason about norms and interact in a normative environment. Thus, we show that coherence-driven normative agents can be put to use in the evolution of the behaviour of agents and of the contents of regulatory systems. Finally, coherence as a motivational criterion is contrasted against other forms of motivational theories of agency.

# Resum

En aquest treball es tracta el problema de la introducció de flexibilitat i adaptabilitat en el disseny d'agents autónoms en el context d'agents situats en entorns regulats. Es discuteix el fet que les arquitectures cognitives, com les basades en la teoria BDI quedin curtes per realitzar el raonament autónom dels agents. Una de les raons importants, ès la carència de motivacions clares per escollir un objectiu o una acció a perseguir. En comptes de la filosofia dirigida per les intencions en una arquitectura BDI, necessitem un formalisme que seleccioni dinàmicament les metes i les intencions considerades com a restriccions entre els elements cognitius. Per tant, l'element central en aquest treball, es refereix a la necessitat que les arquitectures d'agent incorporin un criteri de motivació que pugui ser calculat en termes dels seus elements cognitius, mentre preservi les característiques formals que fan actractives les arquitectures BDI. Aquest treball proposa la teoria cognitiva de la coherència com un criteri de motivació per tal que els agents raonin i prenguin decisions autònomes.

Ens basem en la teoria cognitiva de la coherència proposada per en Paul Thagard. El terme coherència es defineix com la qualitat o l'estat "ser coherent", especialment com a relació entre les parts que sigui lògica, ordenada i estèticament compatible. Un sistema coherent és interdependent, cadascun dels seus fragments d'informació contribueix en la coherència. Considerem la proposta de la coherència d'en Thagard per maximitzar la satisfacció de les restriccions entre els fragments d'informació i, la utilitzem, en el disseny d'agents autònoms. En aquest llibre s'avana l'estat de la qüestió proposant una formalització computacional de la coherència que inclou un mecanisme per calcular el seu valor entre parells de fragments d'informació formalitzant una coherència deductiva.

La part central del llibre proposa una arquitectura basada en la coherència de l'agent que amplia una arquitectura BDI d'acord amb el concepte de coherència. Aquesta arquitectura, preservant les característiques formals de les arquitectures basades en BDI, incorpora coherència com a impuls motivacional central i raona sota incertesa. D'acord amb l'evaluació experimental, demostrem la viabilitat dels agents dirigits per la coherència. S'amplia més enllà l'arquitectura basada en la coherència per a agents normatius, per tal de raonar sobre les normes i la seva interacció en un entorn normatiu. D'aquesta manera, es demostra que els agents normatius dirigits per la coherència es poden utilitzar en l'evolució del comportament dels agents i en l'evolució dels continguts del sistema regulador. Finalment, la coherència com a criteri de motivació es contrasta amb altres teories motivacionals d'agència.

# Acknowledgements

This book is the result of a truly collaborative work towards which a great many people have contributed. I owe my gratitude to all those people who have made this book possible and because of whom my graduate experience has been one that I will cherish forever.

I would like to express my deepest gratitude and appreciation to my advisor Prof. Carles Sierra for his constant support, guidance and great amount of patience. His rare combination of intellectual brilliance, clarity of thought and an ability to get into any level of detail was not only a marvel to watch, but an inspiration to follow. His flexibility and generosity helped me get through the many difficult situations that preceded writing this book. For all that I achieved in this book, I am indebted to him.

This book would not have been possible without the mathematical rigour and attention to detail provided by my co-advisor, Dr. Marco Schorlemmer. I have given him many testing times, yet his insistence on perfection and quality has resulted in a book that is worthwhile. I am also thankful to him for carefully reading and commenting on countless revisions of various parts of this manuscript and for being so understanding and supportive.

I am immensely grateful to Dr. Pilar Dellunde of the Universitat Autonoma de Barcelona (UAB), who was one of the main collaborator in defining the mathematical formulation of deductive coherence and the subsequent proof theoretical analysis (Chapter 3 and 4). I have greatly benefited from her expertise to form a basic understanding of logic and its role in analysing systems. Her support also helped me through the trying times with logic.

I will not be able to express enough gratitude for the collaboration, guidance and support provided by Prof. Henry Prakken of the university of Utrecht during the second half of the book (Chapter 7 and 8). I have not only benefited from the crucial technical expertise needed for defining coherence-driven argumentation, but also from his profound understanding of academic research. I am grateful to him for the time and effort spent in discussions, guidance and above all for being the friend and mentor he was to me.

I would like to acknowledge Dr. Julian Padget of the University of Bath for hosting me at the department and enriching my research experience. I thank Prof. Castelfranchi, Prof. Leon van der Torre, Dr. Guido Boella, and Prof. Dov M Gabbay for the many enriching discussions and insightful comments. I also

thank the Normative Multi-agent Systems research community for providing a continued platform for discussions and collaboration.

I am thankful to Dr. Pablo Noriega for supporting me throughout my graduate years. His insightful comments and constructive criticisms at different stages of my research helped me understand and enrich my ideas. Many a times I relied on him for encouragement and for the lavish confidence bestowed on my ideas. I also wish to thank all the IIIA members, both scientific and administrative for creating a very friendly, open and stimulating environment. I would like to mention Manu, Adrian and Dani for their support with the many procedural and administrative tasks.

The support of many friends within and outside the academic community helped me overcome setbacks and stay focused on my book. I greatly value their friendship and I deeply appreciate their belief in me. I am also grateful to my Indian friends and families for the much needed support in a foreign country. Me gustarìa dar las gracias a Rosa, que me ayudaron con el trabajo de la casa y cuidaba de angelina con mucho amor y dedicación, que fue muy admirable.

Most importantly, I would like to express my deepest gratitude to my parents to whom this book is dedicated. They have been a constant source of love, encouragement, and support during all these years. Their appreciation for science, openness to life and an attitude to embrace challenges have made profound influences in my life. I would also like to express my heart-felt gratitude to my family. This book would not have been possible without their love, understanding and encouragement. I am indebted to my loving husband Rosh, who has been by my side supporting me both technically and otherwise. His criticisms from a reader's perspective have helped me to express my ideas better. He has been very patient and understanding, supporting me lovingly in many bad moments. I am thankful to Joe, Angelina and Rosh who were willing to accept this work as a part of myself. I am also thankful to my dear sisters and brother (to Manju, for helping me with the GUI code for Chapter 6) and to my extended family for all their support, love and encouragement.

Finally, I appreciate the financial support from the OpenKnowledge (http://www.openk.org) Specific Targeted Research Project (STREP), which is funded by the European Commission under contract number FP6-027253 and the Generalitat de Catalunya, under grant 2005-SGR-00093, and the Spanish project "Agreement Technologies" (CONSOLIDER-INGENIO 2010 CSD2007-0022) that funded parts of the research discussed in this book.

# Part I

# Setting the Stage

# Chapter 1

# Introduction

*"Freedom is not worth having if it does not include the freedom to make mistakes."*
Mahatma Gandhi

This book is in the field of autonomous agents and multiagent systems. In this chapter, we give a motivational overview of the field and introduce the research objectives. There are two main research objectives addressed in this book (Section 1.1). The first objective is centered around the ideal of making software agents more autonomous by making them more flexible and adaptive. This looks for reasoning formalisms that incorporate uncertainty and dynamism in the world model without loosing the type of formal qualities that make BDI-like architectures so attractive for testability and reliability reasons. The second research objective addresses application of these autonomous agents to normative multiagent systems, an important motivation for this book. It focuses on ways to make autonomous agents social, capable of reasoning and deliberating about norms and forming sustainable agent societies. In the second section, we highlight the important contributions of this book. The first main contribution is the proposal of coherence-driven agents based on the cognitive theory of coherence as proposed by Paul Thagard [Thagard, 2002]. This includes a coherence framework with a formalisation of *deductive coherence* and a coherence-based architecture with a reasoning algorithm for coherence-driven agents. The second contribution is to model such agents as normative agents that are capable of reasoning about norms and modelling consensus on norm adoption. Finally we outline the organisation of the rest of the book in Section 1.3.

## 1.1 Motivations

Multi-agent systems (MAS) are a well-acknowledged methodology to model complex software systems and simulate intelligent behaviour mainly through interactions between autonomous entities having different information and/or conflicting interests. Research on Agents and MAS has matured during the last decade

and many effective applications of this technology are now being deployed. Distributed healthcare management, e-commerce and e-governance, digital ecosystems, and entertainment and gaming are some of the emerging areas where autonomous agents and MAS are the natural technology of choice.

Some of the characteristic features that are shared by the above mentioned applications are:

1. they are composed of loosely coupled autonomous complex systems
2. they are realised in terms of heterogeneous components and legacy systems
3. they dynamically manage data and resources
4. they are often accessed by remote users and/or in collaboration

For example, a MAS for assisted cognition for elderly patients co-ordinate among various services such as monitoring, providing decision making and warning or reminder services. In its simplest form, such a system would be made up of a series of agents, like monitors and mobile robots capable of reminding, alerting and advising the assisted person. All the actors in the system would clearly be capable of carrying out individual reasoning, but would also need to collectively reason about the situations which can occur [Cesta et al., 2003].

However, the use of MAS at the deployment level is more for providing infrastructures to interoperate between different data formats, integrate different types of services, and unify information gathered from different sources. There is still a lack of technology readiness when it comes to applying MAS consisting of autonomous agents taking independent and autonomous decisions. For example, until recently agents modelling NPCs (Non Player Characters) in virtual worlds and online games [Aranda et al., 2008] have been painstakingly hardcoded by prethinking every potential encounter they might have in the course of the interaction. Fortunately, the situation is changing today and virtual worlds are seen as one of the most potential developing environment for introducing real intelligence in artificial agents. Their "relatively unsophisticated environment" makes it more practical to control and test the autonomous behaviour of artificial agents.

The increasing complexity of such systems and applications not only require that autonomous single agents become more and more intelligent and real, but groups of such agents most likely heterogeneous, interact and share information to achieve their individual goals, while also contributing to the collective goals of the system. For example, agents in mobile health management (providing health services to patients on the move) may need to share information and patient data, health care policy, and information on previous health history. They may also need to take into account rapidly changing national and international laws and regulations concerning the privacy of medical data and the security policies concerning transactions, may need to set up operational norms, and may even need to negotiate on some of the terms based on the specific needs and available services.

Hence in this book we explore two dimensions of agency, a cognitive dimension attempting to accomplish a more flexible and adaptive reasoning capability and a social dimension exploring normative reasoning and interactions in a regulated environment. In particular we try to identify and understand those characteristics that make autonomous agents and MAS suitable for the kind of applications mentioned above. These research problems are formulated in the next subsections.

### 1.1.1 Autonomous Agents

The use of agents making decisions and performing actions in real time while considering the effects of their actions and adapting to dynamic changes in the environment has increased significantly in the context of typical applications of MAS as discussed above. Such agents are alternatively called rational as in Wooldridge et al. [Wooldridge, 2000], autonomous as in [Maes, 1991] or intelligent as in [Russell and Norvig, 2003]. In this book, we use the term autonomous to represent such agents because we concentrate on the capability of the agent to make their decisions and actions without external intervention.

The BDI family of agent models originated from Rao and Georgiff are arguebly some of the most important existing models for designing such agents [Rao and Georgeff, 1995]. A BDI based reasoning process consists of a deliberative cycle in which an agent decides what state of affairs it wants to achieve from among all those desirable states of affairs [Dastani et al., 2003, Shoham, 1993, Rao and Georgeff, 1995]. A main aspect of BDI theory is that it helps selecting what action to perform at each moment. The model focuses on the role of intentions as they constrain the reasoning an agent is required to do in order to perform an action. Once a set of intentions are created and their associated preconditions (in the form of a set of beliefs) are met, then it is immediate that these intentions are realised. BDI models try to reduce the attention problem of an agent by providing an intention to focus on.

However, a key challenge for the BDI family of architectures in general is the need to formalise defeasible (non-monotonic) reasoning, and associated conflict resolution mechanisms. The BOID [Broersen et al., 2002] extension is designed specially for conflict resolution arising between some cognitive elements of an agent and its obligations. The BOID architecture characterises generated candidate goal sets as extensions of a prioritised default logic theory in which rules for inferring goals are modelled as defaults [Reiter, 1987], and a prioritisation of these defaults resolves conflicts between mental attitudes. However, in a BOID architecture, prioritisation on cognitive elements of agents to resolve conflicts is due to different agent types which are identified beforehand. For example, a selfish agent would always prefer goals generated from private desires than those from obligations. And a duty-bound agent would prefer the opposite. This, in our opinion, is not an efficient conflict resolution mechanism because such a mechanism should ideally take into account dynamic changes in a situation and possibly changes in cognitive elements of the agents.

Another way of resolving conflicts or choosing from competing cognitive el-

ements is by introducing preferences as in the graded BDI model (henceforth referred to as *g-BDI*) proposed in [Casali et al., 2005]. The motivation in this work stems from an assumption that an agent's model of the world is incomplete and uncertain. Introduction of degrees is an attempt to capture and represent this uncertainty better in an agent's model. Using a g-BDI model reduces the ambiguity in selecting among the intentions since the degree of an intention is interpreted as its preference or priority and a higher degree implies a higher priority. However, one of the main problems of the BDI family of models is that they follow a linear reasoning structure. That is, an agent choses one or more desires to satisfy and then looks for intentions or plans to realise these desires, thus failing to evaluate desires and intentions in the context of other cognitive elements put together.

Another growing body of work in this context is the literature on argumentative agents that attempts to introduce defeasible reasoning models [Atkinson, 2005a, Amgoud et al., 2000, Modgil, 2008]. An argumentative agent does not reason with basic cognitive elements such as beliefs, desires or intentions, but with arguments computed from these cognitive elements. An action or an intention is selected from a set of intentions based on arguments that support the action. Hence, an action that is supported by the winning argument is chosen as the next action to pursue. Argumentative agents overcome some of the limitations of the BDI family of agents since arguments are generated considering the entire knowledge base of an agent and moreover they are defeasible and hence conflicts among cognitive elements are discovered in the process of constructing arguments that attack or defeat existing arguments.

Most argumentation systems instantiate the general framework of Dung that starts with a set of arguments and binary defeat relations and then determines the set of arguments that can be accepted together [Dung, 1995]. In some of them, tree structured instrumental arguments are composed by chaining the propositional rules with the top of the tree as the high level goal and leaf nodes as primitive actions. A set of instrumental arguments are chosen from sets of conflict-free instrumental arguments that maximise the set of agent goals realised. And some of them further include a preference relation among instrumental arguments based on the value or utility which roughly characterises the worth of the goal and its cost of realisation. A given ordering on values advanced by arguments then determine defeats among arguments [Atkinson, 2005a]. Some of these proposals also include a formal construction of the arguments in an underlying BDI type logic.

One important limitation of argument-based systems is that they tend to be very brittle by demanding conflict-free sets of arguments to be accepted as support for a goal or an action. Whereas in reality, it may only be possible to reduce conflicts but not eliminate them all together. Further, most realisations of argumentation logics only have a binary form of attack relations and are not suitable for modelling uncertainty, though this trend is changing in recent systems. Another limitation that argument-based systems share with BDI-based approaches is that their reasoning progresses in a linear fashion starting from

selecting a goal or a set of goals to realise and then choosing instrumental arguments that support the goals. Alternatively, to resolve conflicts, and more importantly to select among the set of goals, beliefs and intentions, we believe an agent should look at all the relevant information it possess and then should evaluate which subset is more conflict-free from a global perspective.

To summarise, the main limitations of the above approaches are:

- There is a lack of clear cut methods by which some desires are promoted to the level of intentions.
- Even when methods exist, they do not take care of any potential conflicts that exists among desires or among other cognitive elements.
- Most discussed methods are not dynamic in readjusting to new or changed information.
- While the argument-based systems are the most dynamic since they depend on arguments which are constructed on the fly, the values which they use to resolve conflicts are decided a priori.
- None of the methods discussed above select cognitive elements that are most conflict-free from a a global perspective.
- All methods discussed follow a linear reasoning structure starting from a set of beliefs to chose among a set of desires and finally arriving at a set of intentions that realise the set of desires.

Given that, the current state of the art does not fully address the issues we have raised here, we put forward the following research objectives:

> To establish a suitable framework to model autonomous reasoning in agents that can incorporate uncertainty and dynamism in the agent's world model and is capable of resolving conflicts while not loosing the type of formal qualities such as testability and reliability.

This objective may be decomposed into sub-objectives. The first sub-objective is to find a formalism to design an autonomous agent. The idea is to look along the lines of BDI and argumentation logic while overcoming those limitations discussed previously. For example, unlike the intention-driven philosophy in a BDI logic, we need a formalism which would dynamically select intentions based on a global constraint maximisation. The second sub-objective is to define an agent architecture based on the defined formalism. This should further include a reasoning procedure for agents modeled with this formalism. The third sub-objective is to prove that the proposed formalism and architecture when implemented models an autonomous agent with the discussed properties. Concisely, the three sub-objectives are the following:

1. to find a formalism to model autonomous agents that are capable of resolving conflicts under dynamic and uncertain scenarios.

2. to define an agent architecture based on the defined formalism along with an agent reasoning algorithm.

3. to show that the defined architecture models autonomous agents with the specified properties.

### 1.1.2 Autonomous Normative Agents and Normative MAS

An interesting mechanism to co-ordinate the interaction of autonomous agents within a MAS is by making use of norms. Norms while prescribing the accepted behaviour of agents also respect agent autonomy on norm compliance. There is an increasing interest in norm regulated MAS in the computer science community, due to the observation in the AgentLink Roadmap [Luck et al., 2005]—a consensus document on the future of multiagent systems research—that norms must be introduced in agent technology in the medium term for infrastructure for open communities, reasoning in open environments and for trust and reputation. Since then an active community of researchers evolved focusing on norms and normative aspects of MAS. Based on a series of workshops, a consensus evolved as to what can be considered as a norm regulated MAS (referred to as a normative MAS). We quote here one of the definitions most aligned with the perspectives of this book.

> A normative multiagent system is a multiagent system organized by means of mechanisms to represent, communicate, distribute, detect, create, modify, and enforce norms, and mechanisms to deliberate about norms and detect norm violation and fulfilment.

It was remarked in Section 1.1.1 that autonomous agents should be equipped with an effective conflict resolution strategy. This is particularly relevant for autonomous agents situated in a normative MAS (hence forth will be referred to as *autonomous normative agents*) where conflicts among intentions motivated by private goals and those motivated by norm compliance are prevalent. There have been many attempts in the recent past to design agents that could handle such conflicts effectively [Moses and Tennenholtz, 1995, Conte et al., 1999, Boella et al., 2006, Pasquier et al., 2006, López et al., 2002, Kollingbaum and Norman, 2003, Noriega, 1997]. Many of these efforts are focused towards extending the cognitive agent theory (for instance BDI theory) with explicit representation of norms (BOID [Broersen et al., 2002], EMIL [Conte et al., 1999], and NoA [Kollingbaum and Norman, 2003]). However, the kind of conflict resolution strategies employed in most of the current literature limits to prioritising statically among norms and private goals of an agent. That is, a norm priority agent will always prefer norm compliance over satisfaction of private goals when there is a conflict. Hence, it is necessary to extend the features discussed for autonomous agents to autonomous normative agents.

In addition, an autonomous normative agent may need to participate in the set-up or adaptation of norms. This means an agent may need to generate norm proposals, reason about norm proposals of others, and deliberate to reach consensus on norms. In the literature, norm generation and normative agreement are fairly new areas of research and there are no prominent methods so far. However, norm generation is similar to intention generation by an agent that reasons about how to achieve its goals, while normative agreement is similar to reaching agreement on a course of action to solve a problem. For both phenomena logic-based argumentation models have been proposed [Bench-Capon and Prakken, 2006, Amgoud and Prade, 2009] most of which instantiate the general framework of Dung [Dung, 1995]. As discussed in Section 1.1.1, argumentation systems based on Dung's abstract argumentation framework do not take into account uncertainty in the world model of agents and cannot accommodate inconsistency in an accepted set of arguments. Since generating arguments and support for arguments are at the core of a deliberation process to agree on norm proposals, the argumentation system needs to be flexible and expressive.

Hence, the second part of the book deals with autonomous agents and their interaction in a normative MAS. In particular, we care about designing agents that can interact autonomously in a normative MAS by means of an argumentative process deliberate about norms. By this, we emphasize the fact that we not only are concerned with making autonomous normative agents, but are looking at ways to make a normative MAS sustain and adapt over changing situations. As discussed earlier, such agents and systems that adapt are necessary to most MAS applications. Hence, the research objective in the context of autonomous normative agents and normative MAS is the following:

> To design autonomous normative agents and to design a mechanism for such agents to interact and together form sustainable normative MAS.

This can be decomposed into two sub-objectives as follows:

1. To design normative autonomous agents that can
   - reason about norms autonomously,
   - generate norm proposals, and
   - reason about norm proposals of other agents.
2. To design a mechanism for autonomous agents to deliberate about norm change in a normative MAS.

## 1.2 Contributions

The two main contributions of this book are a proposal of coherence-driven agents based on the cognitive theory of coherence [Thagard, 2002] and a

coherence-driven argumentation system for such agents to deliberate about norm adoption. In this section we briefly go over the arguments that make coherence an interesting and suitable theory for the kind of agents and MAS discussed in this book.

### 1.2.1 Autonomous Agents

Some of the properties we would like to have in autonomous agents are the ability to reason taking into account global constraints and the ability to adapt to situational changes (Section 1.1.1). One of the primary factors that facilitate this is a suitable representation of the cognitive elements. In a BDI architecture, cognitive elements are represented in isolation with no explicit reference to possible constraints among them. Such a representation makes it hard to keep track of constraints and more important incorporate them into a reasoning process. Ideally, a goal should be pursued that is not only most desired, but also that has least conflicts with other goals (that may be already in pursuit), other plans and beliefs. A similar process should be followed to incorporate new perceptions and pursue plans. Coherence-driven agents facilitate such a reasoning process with the architecture representing not just cognitive elements but any positive or negative constraints that exist between pairs of elements. Hence, with this architecture, it is possible to maximise satisfaction of constraints at a global level by a process of *maximisation of coherence*. In addition, the effects of dynamic changes in situation are understood by simply updating the cognitive elements in the agent's theory and re-computing satisfaction of constraints.

The coherence-based architecture we propose in this book is inspired by the *theory of coherence*. According to this theory, there are coherence and incoherence relations between *pieces of information* depending on whether they support each other (yielding a positive constraint) or contradict each other (yielding a negative constraint). If two pieces of information are not related, then, there is no coherence (constraint) between them. Based on the characterisation of Thagard, we propose a coherence framework consisting of a *coherence graph* and certain computable functions operating on the graph. A coherence graph consists of nodes to represent the pieces of information and weighted edges to represent constraints between them. Given such a coherence graph, Thagard defines a mechanism to compute the overall coherence of the graph based on maximising constraint satisfaction between pairs of nodes. Certain principles are also defined to characterise and differentiate various types of coherence relations that might exist between pairs of pieces of information. Using the principles of deductive coherence, we define a *deductive coherence function* to compute deductive coherence between pairs of pieces of information of a coherence graph.

We then propose a coherence-based architecture based on the coherence framework. For this, we extend the popular BDI agent architecture with the notion of coherence. By so doing we move away from the intention-driven philosophy of the BDI architecture while retaining the logical properties of the cognitions. Coherence is introduced as the central motivational drive for agents and intentions in a coherence-driven agent are chosen based on the coherence

maximisation of the agent's cognitive elements. Finally, we have evaluated the feasibility of our proposal with empirical analysis and compared it to performance of humans and near optimal algorithms in a restricted setting.

Thus, the main contributions in this book in the field of autonomous agents are the following:

1. Formalisation of a coherence framework based on Thagard's theory of coherence.

2. Definition of a coherence-based agent architecture for autonomous agents consisting of an algorithm for coherence-driven agent reasoning.

3. Empirical evaluation of coherence-driven agents.

### 1.2.2 Autonomous Normative Agents and Normative MAS

The very arguments for coherence to be used in modelling autonomous agents may be extended to the case of autonomous agents with normative capabilities. As argued in Section 1.1.2, conflicts among cognitions are more likely when goals due to norms conflict with personal goals. Due to its representation and global maximisation of constraints, a coherence-based framework lends itself naturally to discovering conflicts. Hence we extend the coherence-based architecture to autonomous normative agents by introducing cognitive elements corresponding to norms in addition to those corresponding to beliefs, desires and intentions. For deliberation on norm adoption, we build upon an argumentation system. We choose argumentation technology since it has emerged as one of the most promising processes for multi agent deliberation with minimal assumptions on the initial positions of the agents, the common knowledge they share, the type of dialogue they engage in, or their motivations [Rahwan et al., 2003b]. In the proposed argumentation system, the notion of an argument consists of a *claim* and its *support* where support is defined in terms of coherence. Since agents are motivated by coherence, it is natural to compute a coherence-driven support. Each agent in a deliberation also evaluates an argument based on a coherence maximisation incorporating the argument into its coherence graph. Unlike traditional argumentation systems, such an argument incorporates degrees of support, and resulting argumentation systems are more tolerant to inconsistencies among arguments.

Thus the main contributions in this book in the field of autonomous normative agents and normative MAS is the following:

1. Definition of a coherence-based architecture for autonomous normative agents

2. Definition of an argumentation system based on coherence for deliberation on norm adoption.

### 1.2.3 Other Contributions

There are two contributions not directly intended nevertheless important in the field of artificial intelligence, cognitive science and economics. The first is the logical formalisation of the cognitive theory of coherence. We have analysed coherence formally, studied its logical properties and proposed a precise computable function to build a coherence graph. This is useful not only to build coherence-driven agents, but also for experiments in physiology and cognitive science and thereby making accessible the use of coherence to a wider audience.

The second contribution is our analysis of coherence in the context of other rationality theories. The game-theoretic concept of Nash equilibrium is one of the better known performance criteria to analyse strategic interactions amongst decision makers [Fudenberg and Tirole, 1991]. However, a number of assumptions make the concept of Nash equilibrium less useful in the context of autonomous agents. Firstly, it is defined for interactions among rational agents where rationality is often interpreted in the neo-classical economic sense of strict utility maximisation. However, strict utility maximisers are just one type of agents and, we need to be able to model different types of agents. Secondly, the concept of Nash equilibrium is developed only for situations where agents have perfect information and common knowledge about the utilities of outcomes of all agents involved. In most cases, these two assumptions do not hold for autonomous agents. A third assumption is that utility maximisation assumes a given ordering of preferences and most often also assumes that this ordering remains static during the interaction. However, a preference ordering of outcomes is a result of maximisation of satisfaction of multiple constraints that exists among an agents cognitive elements. Consequently, a preference ordering should ideally reflect the changes in the knowledge base of an agent, that, unfortunately, can neither be assumed nor remain static.

We in this book prove that coherence maximisation can emulate the properties of a utility maximising function, while getting rid of the strong assumptions that makes utility maximisation less useful. This we see as the first step in having theory of rationality that is more general than the economic notion of strict utility maximisation.

## 1.3 Organisation of the Thesis

This book is organised in four parts discussing each of the four components of the book. Below we give the organisation of these parts into chapters and briefly introduce their contents.

**Part I** contains two chapters including the present chapter which introduces the motivation for this book. Chapter 2 discusses those theories and research findings that serve as the base for the work on this book. Emphasis is given to introducing Thagard's theory of coherence which helps the reader to understand the basic notions of coherence and how it differs from other related theories. It also compares and contrasts the theory of coherence with some of the important

related advances in the field.

**Part II** is organised in four chapters and addresses the first research objective of this book. Two chapters (Chapter 3 and Chapter 4) focus on *finding a formalism to model autonomous agents that are capable of resolving conflicts among cognitions and norms under dynamic and uncertain conditions.* In Chapter 3, we introduce a generic coherence framework, which can be used to create coherence-driven agents. We discuss in this framework how pieces of information can be organised in the form of a graph, along with the necessary computable functions to evaluate and maximise the coherence of such a graph. We then specialise the formulation for a particular type of coherence, namely deductive coherence. We derive a deductive coherence function based on the deduction relation of a logic, however the function we define is independent of the underlying logic. In Chapter 4, we introduce a proof-theoretic characterisation of coherence focusing on deductive coherence. We discuss the formal properties of coherence, and illustrate how these properties help us to derive coherence values between pieces of information.

Chapter 5 focuses on defining an agent architecture based on the coherence framework defined in previous chapters. In particular, we define a coherence-driven agent as a cognitive agent whose utility maximisation is achieved by coherence maximisation. For this purpose we define certain specific graphs corresponding to a cognitive agent. We adapt concepts from multi-context systems so that a coherence-driven agent can reason with its cognitions. We later sketch a procedure an agent may follow in the context of an action selection problem.

In Chapter 6, we prove experimentally the feasibility of a coherence-driven agent and analyse its performance. In particular, we prove the hypothesis that *the performance of a coherence-driven agent is indistinguishable or comparable to the performance of humans and near-optimal algorithms tuned for a specific application.*

Versions of Chapters 3, 4, and 5 have been published in [Joseph et al., 2008b, Joseph et al., 2009b, Joseph et al., 2008a]. Preliminary ideas on Chapter 6 has been published in [Joseph et al., 2010].

**Part III** is organised in three chapters and discusses the second research objective of this book. In Chapter 7, we define an *autonomous normative agent* and discuss an extension of coherence-based architecture to autonomous normative agents. We also focus on norm generation and evaluation aspects of these agents. Chapter 8 proposes an argumentation system for norm deliberation among coherence-driven agents. We focus on a deliberation protocol and the conditions under which coherence-driven agents reach consensus on norms.

Chapter 9 takes a step back to analyse the kind of agents and applications for which a coherence-based model is interesting. We place coherence in the context of other rationality theories and argue in favour of coherence to play a key role in the design of rational agents. In particular, we prove that coherence maximisation can emulate the functionality of other utility maximising functions.

Chapters 7 and 8 have been published in [Joseph and Prakken, 2009]. Chapter 9 has been published in [Joseph et al., 2009a].

**Part IV** concludes the book by discussing the main contributions. We also point to some of the more relevant future work which advances the research initiated in this book. We conclude the book by providing certain insights into the type of applications for which we would like to use coherence-driven agents. We do so by analysing the reasoning of coherence-driven agents in a real-world scenario where a few southern regions of India deliberate on sharing water.

# Chapter 2

# Background

To summarise the introduction, there are two main contributions of this book. The first is the proposal of coherence-driven agents based on the cognitive theory of coherence and the second is the proposal of an argumentation system for coherence-driven agents to deliberate on norm adoption. This chapter presents an account of the relevant motivational, computational, and application background for the research presented in this book. We intend to introduce flexibility and autonomy in agents by taking a new perspective of the theories of agency. Hence, in Section 2.1, we trace some of the prominent motivational theories discussing their relevance in the context of autonomous agent design. We focus on the theory of coherence discussing its salient features as a motivational theory for agent design and decision making. Section 2.2 discusses some of the actual computational realisations of autonomous agents based on the discussed motivational theories. Particular emphasis is given to BDI-based architectures since this book uses a BDI architecture as the base. Section 2.3 gives the background and important research in normative MAS, the primary application area for the research presented in this book. We focus on the issue of norm deliberation both at the single agent and at the multiagent level [Conte, 2001, Conte et al., 1999, Boella et al., 2009]. Section 2.4 is a linkage to different interpretations and other applications where coherence has been used. Concluding remarks are in Section 2.5.

## 2.1 Motivational Theories of Agency

Computational models of motivation are algorithms that cause artificial systems to act. Motivational theories for artificial systems tend to focus on a particular aspect of causation in order to create artificial systems that exhibit some particular form of behaviour. There are primarily three types of causation: biological, psychological and social based on the main causation process involved. In this section, we discuss some of the main motivational theories for agent design and in general for reasoning and decision making.

### 2.1.1 Intentional Stance

Agent theories based on the *intentional stance*, a psychological motivation theory, are the most common ones to model autonomous agents. These are based on *folk physology*, which is often used to predict the behaviour and actions of complex systems like humans [Pitt, 2008]. It is a convenient way to describe our mental states in terms like belief, desire, hunger, pain and so forth. For example

> Anna took the umbrella because she *believed* that it is going to rain.

Here, Anna's behaviour of taking the umbrella can be explained in terms of her attitude towards the fact that it is going to rain. It is philosopher Dennett who first introduced the term *intentional system* to describe entities "whose behaviour can be predicted by the method of attributing certain mentalistic attitudes such as beliefs, desires and rational acumen" [Dennett, 1971]. Some of the important categories of mental states or attitudes identified in the literature are given as below [Wooldridge, 2000].

- Information attitudes — those attitudes an agent has towards the information about its environment. The corresponding mental states are knowledge and belief.
- Pro attitudes — those attitudes an agent has that tend to lead it to perform actions. The corresponding mental states are goals, desires, and intentions.
- Normative attitudes — those attitudes that tend it to behave in a particular way. The corresponding mental states are obligations, permissions and authorisation.

Most formalisations of autonomous cognitive agents choose some of these mental attitudes and study the relationship between them. As discussed in Chapter 1, one of the prominent formalisation following this categorisation is Rao and Georgeff's BDI (Belief, Desire, Intention) logic. Intuitively, an agent's beliefs correspond to information the agent has about the world. They may be incomplete or incorrect. An agent's desires represent states of affairs that the agent would, in an ideal world, wish to be brought about. These desires may be inconsistent. Finally, an agent's intentions represent desires it has committed to achieving. The BDI family of models recognise the primacy of belief, desire, and intention in autonomous behaviour of agents. Particular BDI models that center on claims originally propounded by Michael Bratman about the role of intentions in focusing practical reasoning [Bratman, 1987]. Specifically, Bratman argued that rational agents will tend to focus their practical reasoning on the intentions they have already adopted, and will tend to bypass full consideration of options that conflict with those intentions. In this book, we will be only referring to BDI architectures developed with Bratman Philosophy alternately named as IRMA models (for the "Intelligent Resource-Bounded Machine Architecture") [Bratman et al., 1988].

A BDI-based reasoning process consists of a deliberation cycle in which an agent decides what state of affairs it wants to achieve from

among all those desirable states of affairs [Dastani et al., 2003, Shoham, 1993, Rao and Georgeff, 1995]. The output of the deliberation process is a set of intentions (desires that the agent wants to pursue paired with a 'top-level' plan of action) [Bratman, 1987]. Once the intentions are created and their associated preconditions (in the form of a set of beliefs) are met, it is immediate that these intentions are realised. An overview of BDI-based reasoning process is given in the following:

1. Sense the environment to generate beliefs.
2. If there is no plan,
   choose a desire to pursue,
   find a plan to achieve that desire (usually from a "plan library").
3. Decide on the next action to perform from the plan.
4. Perform that action.
5. Every now and then check that the plan is still valid.

As it should be apparent, there are a few major difficulties with this kind of reasoning. The process of action selection in this approach progresses in a linear fashion. Using a possible world semantics, agents associate a set of belief-accessible worlds for each situation. A subset of desire-accessible worlds are then chosen within each of which a branching future represents the choice of actions available to the agent. One major difficulty with this approach is that the interaction among cognitive elements are in a single direction. For example, the set of desires or intentions do not have any effect on the choice of beliefs. One of the natural ways of revising beliefs by taking into account the feedback from desire and intention cognition cannot be naturally modeled in such systems (e.g. cognitive dissonance [Festinger, 1957]).

The second, and more serious difficulty is that, there are no clear ways to chose between the possible intentions. Further, these intentions or some of the desires driving these intentions may be conflicting. Rao and Georgeff have used the *possible world deliberation* approach to resolve this problem. That is, an agent at each situation uses a probability distribution of the belief-accessible worlds. The agent then chooses sub-worlds of these belief-accessible worlds that it considers are worth pursuing, and associates a payoff value to each path. Using a probability distribution on its belief-accessible worlds and the payoff value with each path in its goal-accessible worlds, the agent determines the best plan(s) of action for different scenarios [Rao and Georgeff, 1995]. However, the semantics of the payoff function is not clear nor is standardised. These difficulties trigger extensions and other theories that are improvements over the basic BDI theory.

### 2.1.2 Utility Maximisation

The BDI family of agents based on the intentional stance and is not designed with social capabilities that help an agent to interact in a society. The motivational

theory of utility maximisation is designed to take into account social aspects of agents. Utility maximisation also overcomes the ambiguity surrounding action selection in BDI-based theories by giving a precise criteria of maximising utility when selecting among possible actions. Utility maximisation comes under psychological motivation theories and in particular within the motivational theory of incentives [Mook, 1987]. The expectancy of being rewarded after some responses forms the basis of incentive. This is extended to the theory of decision making and defines that an agent performs the action whose imagined or expected outcome has the highest utility. The principle of maximisation states that an individual will choose the behavioural response that maximises expected utility. Such agents are usually regarded as rational and have been associated with the economic notion of utility, since available well developed theory of utility maximisation has an economic interpretation of utility.

With this interpretation, utility is also not independently defined but is influenced by actions of other agents in the system. A utility maximiser always tries to maximise its utility assuming that it has a perfect knowledge of utilities of outcomes of all possible courses of actions it can take. Utility maximisers are often defined in a social context since more often an outcome depends not only on the choice of actions by a single agent but that of many agents acting together. Hence it is desirable for utility maximisers to take into account the reasoning or the utilities of other agents. This makes it the case that each agent ideally has a model of other agents involved in influencing an outcome.

A second assumption is that, utility maximisation assumes a given ordering of preferences and most often also assumes that this ordering remains static during the interaction. However, a preference ordering of outcomes is a result of maximisation of satisfaction of multiple constraints that exists among an agents cognitive elements (for example the agents beliefs or intentions). Consequently, a preference ordering should ideally reflect the changes in the knowledge base of an agent, that, unfortunately, can neither be assumed nor remain static. These strong assumptions make it a less practical motivational theory for modelling autonomous artificial agents.

### 2.1.3 Reinforcement Learning

If utility maximisation assumes a given order of preferences, another motivational theory of reinforcement learning tries to understand implicit preferences through the mechanism of reinforcement. Reinforcement learning represents a psychological motivation where rewards are used to enforce a behaviour [Sutton and Barto, 1998]. Positive or negative rewards act as a motivational force for taking a particular course of action. In effect, agents learn a function which represents the value of taking a given action in a given state with respect to some task. An agent is situated in its environment and receives perception and outputs action. On each step of interaction with the environment, the agent receives an input that contains some indication of the current state of the environment. The agent then chooses an action as output. The action changes the state of the environment and the value of this state transition is

communicated to the agent through a scalar reinforcement signal. The agents behaviour should choose actions that tend to increase the long-run sum of values of the reinforcement signal. This behaviour is learnt over time by systematic trial and error. Reinforcement learning assumes that the reinforcement stimuli are always provided by a source external to the agent.

In the case of reinforcement learning, an agent does not generate its own motivations or goals to achieve, but only behaves according to the reinforcement training received. There are further developments that help an agent to bootstrap by statically assigning certain general features with a high value of reinforced signal. However, the research on reinforcement learning gives more emphasis to the learning algorithm than to the motivational theory. Another limiting factor is that most learning agents operate with a low level perception of their environment. In other words they are not self-aware nor have a high level perception of goals they are given to pursue. Hence, their capacities to adapt their goals to situational changes are limited.

### 2.1.4 Theory of Coherence

After reviewing some of the main motivational theories used up to now for modelling autonomous agents, we now introduce the theory of coherence as a motivational theory for designing autonomous agents. We will see during the discussion that this motivational theory along with the coherence-based architecture introduced in this book overcomes some of the main difficulties found in other motivational theories. The theory of coherence is a psychologically motivated theory which tries to give an intrinsic, domain independent motivation to an agent. As a motivational theory, it is very general and hence can be used to direct agent behaviour at a high level. Since this theory forms the basis of this book, we discuss the characteristics of this theory in detail and compare and contrast it with other decision theories and philosophical theories. Even though the theory of coherence has been in existence for long, we here refer to Thagard's interpretation of the theory as it is Thagard who proposed a computational formalisation of the theory of coherence.

Thagard postulates that the theory of coherence is a cognitive theory with foundations in philosophy that approaches problems in terms of the satisfaction of multiple constraints within networks of highly interconnected elements [Thagard, 2002, Thagard, 2006]. At the interpretation level, Thagard's theory of coherence is the study of associations, that is, how a piece of information influences another and how best different pieces of information can fit together. Each piece of information imposes constraints on others, the constraints being positive or negative. Positive constraints strengthen pieces of information, thereby increasing coherence, while negative constraints weaken them, thereby increasing incoherence. Hence, a coherence problem is to put together those pieces of information that have a positive constraint between them, while separating those having a negative constraint. Coherence is maximised if we obtain such a partition of information where a maximum number of constraints is satisfied.

The basic concepts in the formalisation of Thagard are that of a set of pieces of information that are represented as nodes in a graph with weighted links, or constraints, between these nodes. Further, some of these constraints are positive (representing coherence) and others negative (representing incoherence), and associated with each constraint is a number that indicates the weight of the constraint. Given these, maximising coherence is formulated as the problem of partitioning the set of nodes into two sets, $\mathcal{A}$ (the accepted nodes) and $\mathcal{R}$ (the rejected nodes), in a way that maximises compliance with the following two coherence conditions:

- if edge $\{v, w\}$ is positive, then $v \in \mathcal{A}$ if and only if $w \in \mathcal{A}$.
- if edge $\{v, w\}$ is negative, then $v \in \mathcal{A}$ if and only if $w \in \mathcal{R}$.

If an edge complies with one of the above conditions, then, Thagard defines it as a satisfied constraint. The coherence problem is thus simply to find a partition that maximises the sum of the weights (called the *strength* of the partition) of the satisfied constraints.

Thagard further proposes six main kinds of coherence: *explanatory, deductive, conceptual, analogical, perceptual, and deliberative*, each with its own array of elements and constraints. Once these elements and constraints are specified, then those algorithms that solve the general coherence problem can be used to compute coherence in ways that apply to specific domain problems.

Thagard has also experimented with many computational implementations of coherence. ECHO is a computational model of explanatory coherence which uses a connectionist algorithm [Thagard, 2002]. Though there is no guarantee that such neural network models for coherence would converge to a coherence-maximising partition, he claims that on small networks it has been shown to give good results.

Thus, Thagard proposes the first major concrete account of coherence that takes us from the abstract notion of coherence to a computational phenomenon that can be evaluated. One of the major shortcomings of his formalisation is that he stops with giving certain principles about calculating values of coherence constraints for different types of coherence. For example, to compute the explanatory coherence value between two propositions say "Mary took the Umbrella" and "It is raining outside", we need to have concrete functions which reflect the underlying explanatory relationship between the two propositions. Thagard in his formalisation proposes some of the properties such a function should have, nevertheless does not go further on it. Without these coherence functions, computable formalisations are not possible. These functions can be thought of as the core of the coherence formalisation. In that sense, the coherence framework we propose in this book attempts to fill this gap by taking the deductive coherence principles of Thagard and defining a deductive coherence function exploiting the deductive relationship between propositions.

### 2.1.5 Comparison with Other Decision Theories

Keeping Thagard's approach to coherence as maximising constraint satisfaction, we try to understand the main concept behind this theory. We associate coherence with an ever-changing system where coherence is the only property that is preserved, while everything around it changes. In cognitive terms, this would mean that there are no beliefs nor other cognitive elements that are taken for granted or fixed forever. Everything can be changed and may be changed to keep coherence. We humans tend to revise or re-evaluate adherence to social norms, our plans, goals and even beliefs when we are faced with incoherence. However, we do not suppose that taking decisions based on coherence imply an unstable system. Our claim is based on the fact that some beliefs are more fundamental than others. Revision of such fundamental beliefs is less frequent compared to other beliefs. In coherence terms, these beliefs are fundamental because they support and get support from most other cognitive elements and hence are in positive coherence with them. Hence, such beliefs will almost always be part of the chosen set while maximising coherence. Similar is the case with desires and intentions while the process of coherence maximisation further helps resolve conflicts by selecting among the best alternatives.

When applied to decision making, this means that we may not only select the set of actions to be performed to achieve certain fixed goals, but also look for the best set of goals to be pursued. Further, since coherence affects everything from beliefs to goals and actions, it may happen that beliefs contradicting a decision made are discarded. There are psychological theories such as cognitive dissonance [Festinger, 1957] that explain this phenomenon as an attempt to justify the action chosen. Thus, with coherence we are looking at a more dynamic model of cognitions where one picks and chooses goals, actions and even beliefs to fit a grand plan of maximising coherence. In concrete terms, a highly desired state of the world (desired in a classical sense) may get discarded in front of a less desired state of the world because it is incoherent with the rest of the beliefs, desires or intentions.

As discussed in [Thagard, 2002], this view of decision making is very different from those of classical decision-making theories where the notion of *preference* is atomic and there is no conceptual understanding of how preferences can be formed. In contrast, coherence-based decision making tries to understand and evaluate these preferences from the available complex network of constraints. The assumption here is more basic because the only knowledge available to us are the various interacting constraints between pieces of information.

### 2.1.6 In the Context of Philosophical Theories

Seen in a broader context, the theory of coherence can draw parallels with other established theories. The philosophers of science have long argued about what "claims" in a theory can be supported. Popper's view on the progress of knowledge [K.Popper, 1962] sees falsifiability as the main driving force, and knowledge as an evolving body that follows a process in which a number of theories

'compete' to account for a problem situation. When a set of theories is set, falsification is then the process that makes some theories fail, while allowing others to survive. In his view survival does not mean truth but 'fitness' to the situation. The notion of truth-likeness is for Popper a notion of verosimilitude ($V(a) = T(a) - F(a)$) that accounts for the comparison between the truth content of theory $a$ and the falsity content of $a$, which permits to rank theories. This concept is similar to the notion of 'strength' of a partition in a coherence graph. Falsification of a theory can be associated to the introduction of a highly incoherent fact that will make certain statements to be removed from the accepted set of claims. Although Popper would reject a complete theory as soon as empirical evidence would go against it, Kuhn [Kuhn, 1962] would consider that scientists tolerate a certain level of anomalies (in our context a certain level of incoherence) for a long time until a revolution happens in which a complete new theory is accepted and an old one rejected. This latter phenomenon may be reproduced in our context, by the fact that partitions in graphs can change abruptly when two theories are similarly coherent and a new experimental result is added leading to a swap in the set of accepted claims. The reconciliation point made by Lakatos [Lakatos, 1976] would be that scientific theories contain a hard core that contains the most crucial claims of the theory plus a protective belt of auxiliary hypothesis that in case of contradiction with the facts will be modified or removed while keeping the central core, of course until a major difficulty is found that leads to a drastic change of the core. As will be apparent in the course of the book, the use of degrees in claims and the algorithmic introduced in the formalisation of coherence shows that we might implement a similar mechanism by eliminating first the auxiliary hypothesis (those with lower degrees of belief) before removing the hard core ones (with higher degrees).

Thus, we see that coherence while overcoming most of the limitations encountered in the discussed motivational theories for agent design, is also in alignment with prominent philosophical theories. This provides us sufficient grounds to incorporate coherence as a motivational theory for agent design.

## 2.2 Computational Formalisms

We now discuss computational realisations of agent architectures based on some of the motivational theories. We discuss the Soar architecture, BDI-based architectures and finally a BDI architecture extended with coherence [Pasquier et al., 2006].

### 2.2.1 The Soar Architecture

Unlike BDI architectures that are based on the well understood BDI theory and Bratman's theory, Soar is primarily an architecture without theories of cognition [Laird et al., 1987, Laird et al., 1991]. Soar is a symbolic cognitive architecture where behaviour is understood as a combination of architecture and content. There are no single motivational theories behind Soar, but an integra-

tion of many philosophies to create intelligent behaviour. The essence of the Soar architecture is a goal directed behaviour with a state space representation. The architecture supports multi-method problem solving, representation and use of multiple knowledge forms, and interaction with the outside world. This is based on the assumptions that cognitive behaviour is goal oriented, reactive, requires the use of symbols, requires learning and operates within a problem space.

An abstract mapping may be made between BDI theory elements and the Soar architecture with intentions being mapped to selected operators; beliefs being included in the current state; and desires being mapped to goals. Bratman's insights about the use of commitments in plans are applicable in Soar as well. For instance, in Soar, a selected operator (commitment) constrains the new operators (options) that the agent is willing to consider. In particular, the operator constrains the problem-space that is selected in its subgoal.

In Soar, the reasoning proceeds with agents sensing the environment, representing perceptions in the working memory, generating all goals that can fire and making a selection among the generated goals with the help of preferences. As is evident, the decision making progresses in the traditional linear structure considering one goal at a time and trying to achieve that goal. If more than one goal is active, then there are ways to introduce preferences among goals such as learning from past behaviour to select among goals. This style of reasoning suffers from the kind of problems of the BDI theory based architectures (see Section 2.1). Further, Soar has additional difficulties in knowledge creation due to its dependency on production rules. To compensate for this difficulty, Soar has incorporated learning as a basic functionality within the architecture, the results of which are encoded as production rules. However, the learning in Soar is not developed sufficiently to compensate for the drawback.

### 2.2.2 BDI family of Architectures

The BDI theory served as the base on which many architectures were proposed and extended to accommodate and adjust the theory to real world requirements. From BDICTL of Rao and Georgeff [Rao and Georgeff, 1991], LORA (the logic of rational agents) [Wooldridge, 2000] to BOID [Broersen et al., 2002, Broersen et al., 2001], there have been numerous proposals to incorporate various levels of autonomy in agent design. We discuss the BOID architecture as a representative architecture particularly because the extension of BDI with norms is interesting for us in the context of normative autonomous agents. Another specific extension we discuss here is the extension of BDI with degrees [Casali et al., 2005].

#### The BOID Architecture

As mentioned in Chapter 1, one of the agent architectures that is particularly designed for agent reasoning under conflict is the BOID (Belief, Desire, Intention, Obligation) architecture proposed in [Broersen et al., 2002, Broersen et al., 2001]. In BOID, there are at least four mental states of beliefs,

desires, intentions and obligations, the architecture having the corresponding four components. These represent conditional mental attitudes since they output beliefs, desires intentions and obligations only for certain inputs. Conflicts are considered between various mental attitudes and resolved based on the ordering of the generation of the outputs. The ordering in turn is based on the generalisation of the already existing categorisation of agents as realistic, selfish, social and stable. That is, an agent designed with an ordering $B > O > I > D$, represented as BOID (note that the naming represent the ordering relation between mental attitudes), will be realistic social and stable. It is a realistic agent since it resolves any conflicts between beliefs and any other mental attitudes in favour of the former. It is a social agent since it places obligation before other planned intentions.

To see how reasoning proceeds in a BOID agent, we describe an agent with the above ordering. Upon receiving an input in the form of belief, desire, intention or obligation, a realistic BOID agent starts with the observations and calculates a belief extension by iteratively applying belief rules. When no belief rule is applicable anymore, then either the O, the I, or the D component is chosen from which one applicable rule is selected and applied. When a rule from a chosen component is applied successfully, the belief component is attended again and belief rules are applied. If there is no rule from the chosen component applicable, then another component is chosen again. If there is no rule from any of the components applicable, then the process terminates —a fixed point is reached— and the output represent one of the mental attitudes.

This model simplifies conflict resolution by identifying various types of agents with each agent having a static preference ordering of mental attitudes and processes any observation in the light of this ordering. However, as we know, conflict resolution under dynamic and uncertain circumstances is a much more complex phenomenon which requires a more careful treatment. However, this was some of the first developments in agent reasoning in the context of conflicts and hence inspired further models to incorporate uncertainty and preference information as a dynamic aspect of the model. One such example we discuss next is the g-BDI architecture. Other argumentation-based architectures which we will discuss in Section 2.3 are improvements over the BOID in that they try to resolve conflicts dynamically.

### The g-BDI Architecture

Since we use the g-BDI along with the multi-context system architecture as the base for the coherence-based architecture introduced in this book, we discuss the essential building blocks of this architecture and its merits while also pondering on possible improvements.

The g-BDI model uses a multi-context system (MCS) architecture originally proposed by Giunchiglia [Giunchiglia and Giunchiglia, 1993, Giunchiglia and Serafini, 1994]. In the work of Casali et al., the MCS specification of an agent contains three basic components: units or contexts, logics, and bridge rules that channel the propagation of consequences between theo-

ries. Contexts in a multi-context BDI are the contexts of belief, desire, and intention cognitions. The deduction mechanism of MCS is based on two kinds of inference rules, internal rule inside each context, and bridge rules between contexts. Internal rules allow an agent to draw consequences within a context, while bridge rules allow to embed results from one context into another [Giunchiglia and Giunchiglia, 1993, Giunchiglia and Serafini, 1994].

In a g-BDI architecture, a degree on a belief represents the confidence on the belief, a degree on a desire represents the preference on the desire while a degree on an intention represents the trade-of between the benefit and cost in reaching a goal by the execution of the intention. In this architecture, degrees are used to select between competing goals and once again to select between competing intentions for the selected goal. Once an intention is selected, then a plan to realise that intention is chosen to be executed.

The architecture uses a many-valued modal logic proposed by Hajek to model uncertainty in different mental contexts [Hájek, 1998]. Hajek developed an approach for uncertainty modelling by defining suitable model theories over suitable many-valued logics. This proposal allows to use well-founded logical frameworks (as diverse many-valued logics) to represent different uncertainty models by adding the adequate axiomatics for each case. The basic intuition behind this approach is to consider for example, the belief degree of a crisp proposition as the truth-degree of a fuzzy(modal) proposition. That is, if belief degrees are modelled as probabilities, then for each crisp formula $\varphi$, the corresponding modal form $B\varphi$ can be considered as a fuzzy formula by associating the truth-value of $B\varphi$ with the probability of $\varphi$. Hence, we can import the axioms of probability theory as logical axioms for modal formulae of the type $B\varphi$.

Once the language of fuzzy propositions are defined such as $B\varphi$, ($B$ being a modality representing probable, necessary, desirable, etc.) and $\varphi$ a crisp proposition, then we can write theories about say, $B\varphi$ formulae over a particular fuzzy logic. This framework makes the selection of an underlying many-valued logic independent to model the uncertainty degree of the fuzzy formula. The particular choice of logic and methodology are discussed in Chapters 3 and 5 as and when we use or extend this framework. Thus, this is one of the important extensions of BDI models of rao and Georgiff and the base for the coherence-based architecture.

However, this approach does not fully meet the flexibility and adaptability requirements we seek for an agent architecture. We list some of the shortcomings of this architecture with respect to the flexibility and adaptability requirements:

1. In this architecture, a desire with the highest degree is always chosen to be pursued. However, this desire may be in conflict with some of the beliefs or other desires already in pursuit. Hence, it is not clear that a desire with the highest degree is always the best to be pursued.

2. In general, the model is not particularly suited to agent decision making in the context of conflicts among mental attitudes and external commitments, since there are no explicit mechanisms to handle or resolve conflicts.

3. The cost of realising a desire is not taken into account while associating a preference on the desire. And the question of feasibility of a desire does not arise until it is chosen and the plans are evaluated to realise it.

### 2.2.3 A BDI Architecture with Coherence

Based on Thagard's formulation of coherence as maximising satisfaction of constraints, Pasquier et al. initiated the application of the theory of coherence in developing a model of agent communication pragmatics [Pasquier et al., 2006, Pasquier and Chaib-draa, 2003, Pasquier et al., 2004]. Based on Thagard's conceptual framework, they developed a theory of the cognitive aspects of a communicating agent, answering such questions as *"When should an agent take a dialogue initiative, on which subject, with whom and why?", "When to stop a dialogue or if not, how to pursue it ?", "How to define and measure the utility of a conversation?" and "What are the impacts of the dialogue on agents' attitudes"*. Although, this work does not propose a formalism for the theory of coherence, it is one of the major initiatives to use coherence theory in the context of multiagent systems. It is also interesting for us, since they use a BDI-based agent architecture to validate the theory. Finally, since agent communication in their context also gives rise to agents taking commitments, their work can also be viewed as agent reasoning in the context of conflicts arising from social commitments and individual preferences. We first briefly discuss their formulation of agent reasoning enriched with coherence and discuss its relevance to the theme in this book.

The coherence framework is implemented as a layer above the belief, desire and intention layer in a BDI architecture. The communication between the BDI base layer and the new coherence layer is two-way. First, the coherence layer receives intentions from the BDI base layer. The work assumes that intentions are generated following a usual BDI reasoning process. Since the BDI architecture does not treat commitments in any special manner, and since here the agent communication is primarily about commitments, it is assumed that the intentions that are received from the BDI base layer are either *social individual intentions*, which concern goals that require some social aspect to be worked on, or *failed individual intentions*, which are intentions with a failed plan or with which no plan is associated.

An agent's state is characterised by

1. sets of perceptions, beliefs and individual intentions, and a set of agent's agenda, that stores all the social commitments from which the agent is either the debtor or the creditor;

2. sets of positive or negative constraints over pairs of elements of the sets above such that every pair belongs to either of these sets;

3. the accepted and the rejected sets to either of which every element of the sets above belong.

The coherence is computed using the formulation of Thagard, however uses a naive but simple greedy algorithm that computes the coherence gain due to flips (between accepted and rejected sets) of each element. The flip corresponding to the maximum coherence gain is selected as the next action (There is actually a utility measure calculated which includes the cost of flipping). If it involves a communication, then the agent takes the initiative to start a dialogue.

Similar to the formalisation of Thagard, this work also does not propose any precise functions to compute coherence values between pairs of elements. Positive or negative constraints between intentions and commitments are determined based on the following rule. "An accepted commitment is the counter part of an intention, commitments in action are the counter part of *intention to* and propositional commitments are the counter parts of *intentions that*. Constraints between the intentional private layer and social commitment layer is inferred from the above relationships or links as well as other logical relationships between intentions and social commitments. Since the coherence layer consists of intentions passed onto from the BDI layer, the only coherence relations that are studied and considered for coherence maximisation are those between internal intentions and social commitments.

In this work, coherence maximisation is applied only to choose between intentions while the rest of the reasoning follows a linear structure. This in our understanding deviates from the very intuition of the theory of coherence since, with coherence, decision making is a dynamic process where perceptions, beliefs, goals, and intentions are selected based on coherence-maximisation. With this approach, constraints may be only revealed partially, and a reasoning process may use certain chains of deductions which links together some beliefs, desires and intentions. It is precisely those negative constraints that are likely to be left out. Finding conflicts or incoherences at the intention level suffers from the same limitations we have discussed for the BDI family of architectures. Further, without a precise computational mechanism to construct coherence graphs, the results of coherence maximisation cannot have useful interpretations.

We in this book aim to overcome these limitations and improve the coherence framework by making it precise, general and formally grounded. However, the most important improvement we introduce here is to position coherence as a fundamental property of the mental states, not just as a tool for resolving conflicts.

## 2.3 Applications in normative MAS

In this section, we take a few representative research from the normative MAS literature for discussion. One of the basic motivations for normative MAS is the assumption that norms help agents to form certain behaviour expectations of other interacting agents while still maintaining assumptions on autonomous behaviour. In this sense normative MAS provides a very promising model for interaction and co-ordination in MAS [Boella et al., 2006]. One of the early introductions of norms in multi agent co-ordination is the work on artificial social systems

by Tennenholtz [Shoham and Tennenholtz, 1995, Moses and Tennenholtz, 1995, Fitoussi and Tennenholtz, 2000]. The problem studied in artificial social systems is the design and emergence of social laws. Shoham and Tennenholtz studied artificial social systems using notions of game theory. Continuing their work, there have been much research in normative MAS both from the social and from the cognitive perspectives [Castelfranchi et al., 2000, Conte et al., 1999, López et al., 2002]. As the work in this book covers both cognitive aspects of normative reasoning by single agents and social aspects of norm deliberation by groups of agents, we analyse the state of the art in both these aspects.

### 2.3.1 Normative Reasoning

The work by Guido et al. provides a comprehensive account of the situations faced by different types of agents in which they could possibly violate norms [Boella and van der Torre, 2004]. Situations include those when there are contradictions between goals and obligations, when violation is preferred to possible sanction, when norm consequences are not understood, or when norm compliance is impossible. The work then proceeds to formalise some of these situations. It does not however address the reasoning of an agent placed in these situations. Thus the work is normative in defining conditions under which norm violation happens more from the perspective of the normative environment than that of an agent. Interestingly, all situations listed in this work are situations where agents suffer certain lack of coherence and a coherence-driven agent can be used to model agents in those situations. In that sense our work and the work of Guido et al. is complementary.

The work of Conte et al. treats norms from the cognitive perspective of individual agents. They claim that some of the most important issues surrounding the study of norms are reasoning about acquiring new norms and norm conformity [Conte et al., 1999, Conte, 2001]. In their work they address the issue of autonomous norm acceptance in agents and how that is instrumental to distributed norm formation and norm conformity in an agent society. The authors describe autonomous norm acceptance as a two step process, first recognising the norm issued by an external entity as a norm, and second, deciding to conform to it. The first step according to the authors would form a normative belief, and the second step would create a normative goal or intention. To move from normative belief to normative conformity, an agent would additionally require the existence of other private goals which would benefit from the normative goal. That is, an agent complies with a norm when the norm is instrumental to solving a private goal. In the presence of such a norm, it is immediate that the normative intention is created and normative conformity is established. This is similar to a BDI reasoning process where once an intention is created, it is immediate that the intention is realised. Obviously, in the presence of conflicts with other private goals, this method does not provide capabilities of autonomous reasoning.

On the contrary, for an autonomous agent to accept a norm, the agent has to understand what a norm really means and its implications in terms of its own cognitions. And to conform to a norm it should know what actions or beliefs are

permitted, prohibited or obliged. In this sense even though their work addresses many relevant issues around normative reasoning, does not provide mechanisms for agents to autonomously reason about norms.

The NoA architecture is specifically constructed for the implementation of norm-governed agents and includes mechanisms that are important for maintaining the normative state of an agent and for the adoption of norms [Kollingbaum and Norman, 2003]. The work addresses circumstances under which it is appropriate for an agent to adopt a new norm, the effects of norm adoption on agent's normative state and whether a newly adopted norm is consistent with the norms currently held by the agent. The NoA architecture has an expressive representation of plans and norms representing explicitly all effects of a plan based on an assumption that *a plan may have multiple effects*. However, it is not clear whether all effects of a plan can be statically determined. A second assumption is that NoA agents are motivated by norms. This, though works well for norm conforming agents, is a very strong assumption that works against the autonomy of an agent. The following rules further clarify this position.

- Implicit-permission-assumption: If an agent has the capability to perform an action then it has an implicit (or default) permission to do so.

- A newly adopted explicit prohibition will override and explicitly restrict, partially or completely, the agent's default freedom (or implicit permission).

Hence, a motivation generated by an obligation or prohibition has been statically assigned a higher priority compared to motivations generated from permissions. Similarly other static rules are used to resolve inconsistencies among overlapping or conflicting norms. Even though an NoA agent exhibits a stable behaviour and resolves conflicts, there is much lack in flexibility and adaptation characteristics.

### 2.3.2 Multiagent Norm Deliberation

In this book, we propose a coherence-driven argumentation for deliberation on norm adoption. This section discusses the basic notions on argumentation and refer to some of the important works in the field. We also give a brief comparison between the general argumentation approach and the coherence-based argumentation, while a detailed comparison can be found in Chapter 8.

In recent years, a growing body of work has been proposed on argumentation approaches to negotiation, persuasion, and in general to agree on a course of action to solve a problem [Atkinson, 2005b, Dung, 1995, Pollock, 1975, Amgoud et al., 2008]. The field dates back to the time of John Pollock, whose work is still part of the state of the art in argumentation with respect to incorporating preferences [Pollock, 1975], and Raymond Reiter who was one of the founders of non-monotonic logic [Reiter, 1987]. However, it is Phan Minh Dung in 1995, who introduced an abstract argumentation framework for

argumentation-based inference which later became a standard framework of reference for all argumentation systems [Dung, 1995]. Dung's framework assumes as input nothing else but a set of arguments ordered by a binary relation of attack. Dung made no assumptions on the structure of the arguments themselves nor on the nature of the attack relations thus enabling instantiation of the framework by various logical formalisms. This framework provided a general and intuitive semantics for the consequence notions of argumentation logic (and for non-monotonic logic in general). It enabled a precise comparison between different systems (by translating them into the abstract framework) and it made the general study of formal properties possible as instantiations of Dung's general theory [Prakken, 2009].

An argumentation framework may be concerned with logics for argumentation or protocols for argumentation or both. The former defines which conclusions can be drawn from a given body of information, while the latter regulates how such a body of information can be constructed in dialogue. Here we concern ourselves only with logic for argumentation. Protocols for argumentation will be discussed in Chapter 8. Even though norm generation has been a less studied topic, several argument-based logics for intention generation have been proposed. As norm generation and intention generation essentially follow similar rules even though stem from different motivations, we use the research on intention generation for the purpose of comparison. Bench-Capon et al. aims to formalise the reasoning model underlying [Atkinson, 2005b]'s dialogue model for disputes over action [Bench-Capon and Prakken, 2006], while Amgoud et al. proposes an alternative account [Amgoud and Prade, 2009]. The essential ingredient in both approaches consists of rules for constructing arguments. These constrain the autonomy of the agents as opposed to the idea of coherence-based argumentation which does not pre-conclude on any winning arguments. Bench-Capon et al. then applies Prakken's accrual mechanism to aggregate arguments for or against the same intentions [Prakken, 2005b], while Amgoud et al. leave the aggregation of such arguments outside the logic and model them decision-theoretically.

The logics of [Bench-Capon and Prakken, 2006, Amgoud and Prade, 2009] instantiate the general framework of Dung, which starts from a set of arguments with a binary defeat relation and then determines which sets of arguments can be accepted together. This is similar to determining partitions of a coherence graph, but in approaches that instantiate Dung's format, support and defeat relations between arguments and the acceptability of arguments cannot be a matter of numerical degree, while sets of acceptable arguments cannot contain conflicts. As will be evident in Chapter 8, on all these points a coherence approach is meant to provide more flexibility, since in reality support, attack and acceptability are often a matter of degree. As you will see in Chapter 8, one possible benefit of this may be a natural modelling of accrual of arguments for the same conclusion. The notion of accrual of arguments is based on the intuitive idea that having more reasons or arguments for a given conclusion makes such a conclusion more credible. By contrast, modelling accrual of arguments is not a simple issue, and research on argumentation has identified different principles that should hold for

performing accrual of arguments in a sound way. In [Amgoud and Prade, 2009] accrual is modelled outside the logic while the logical accrual mechanism of [Bench-Capon and Prakken, 2006] is quite complex.

On the other hand, a strong point of argument-based approaches is that they yield explicit reasons why an outcome should be adopted or rejected, while coherence-based approaches are often criticised for their lack of transparency. The proposal in this book is specifically designed to address this criticism by constructing a coherence graph grounded in logical deduction, thus making explicit why two pieces of information are positively or negatively coherent. As will be evident, this feature is further exploited in defining the notion of argument.

## 2.4 Other Applications

Before concluding this section, we would like to analyse two other applications of coherence one in the area of natural language analysis and the other in the area of legal reasoning. They are interesting both for their formalisation and for their novel applications.

### 2.4.1 Coherence and Linguistic Analysis

In this section, we analyse those proposals that formalise a general notion of coherence but do not follow Thagard's interpretation of coherence as maximising satisfaction of constraints. The theory of coherence has been studied in philosophy, computer science and law, however there are very few attempts to formalise coherence so that it could be used as a general framework. Two of the few such attempts in the field of linguistic coherence are analysed here. Both these works concentrate on linguistic coherence which is the property of a text or conversation being semantically meaningful. However, from the formal perspective, there are overlaps as the principles of coherence essentially stay the same. We compare and contrast their proposals and the work in this book.

The work of Piwek attempts to model dialogue coherence in terms of generative systems based on natural deduction [Piwek, 2007]. The main argument in his work is that it is possible to generate coherent dialogues by relying on entailment relations in an agent's knowledge base. The work primarily deals with information seeking dialogue where the definition of whether an agent knows a fact is equated to whether it can be logically entailed. This is an interesting way to look at dialogue coherence where the concern is semantic rather than structural. However, the properties of cognitive coherence as a relation are neither exploited nor modeled. Coherence in his work refers to the meaning of coherence in a linguistic sense; i.e, *what makes a text or conversation semantically meaningful* whereas the coherence we deal with is a property of the cognitive state. Though coherence is related to entailment, coherence is not equivalent to it, and it is important to capture and model the differences.

The work of Sansonnet et al. models agent dialogue based on the theory of dissonance [Sansonnet and Valencia, 2003]. The theory of cognitive dissonance

states that contradicting cognitions serve as a driving force that compels the mind to acquire or invent new thoughts or beliefs, or to modify existing beliefs, so as to reduce the amount of dissonance (conflict) between cognitions. Their work exploits the drive to reduce dissonance as a cause to initiate a dialogue and later to terminate when this dissonance no longer persists. It is curious to note that many authors who have used the theory of dissonance in dialogue initiation and termination have not considered the possibility that, not all incoherences are dissonant [Pasquier et al., 2006, Sansonnet and Valencia, 2003], but dissonance seeks out specialised information or actions. The most important difference between the work of Sansonnet et al. and the work in this book is that, for them coherence (or the lack of it) is a local phenomena concerning only the new arriving fact and the fact it contradicts with, whereas for us coherence is a global phenomena affecting the entire knowledge base of the agent. As in the case of Piwek, the authors equate coherence with logical entailment.

### 2.4.2 Coherence and Legal Reasoning

Coherence models have also been earlier applied to legal reasoning by Thagard [Thagard, 2004], Amaya [Amaya, 2009] and Bench-Capon et al. [Bench-Capon and Sartor, 2001]. Thagard and Amaya use explanatory coherence to model scenario-based reasoning about evidence, while Bench-Capon et al. use a coherence model in their theory formation approach to case-based reasoning. Amaya tries to apply a notion of coherence in legal justification and studies how notions of fairness and coherence are related [Amaya, 2007]. The work also claims that coherence considerations need to be taken while putting forward an argument along with truth and fairness considerations. In her work, Amaya analyses Thagard's models of coherence as constraint satisfaction and argues that such models should be used in conducting argument justification in legal reasoning. She has analysed different aspects of coherence and has formalised systems of coherence thoroughly. Her treatment clarifies many conceptual issues about coherence. However, apart from suggesting and justifying why coherence needs to be used in legal reasoning, she does not propose a formalisation herself. Finally, the work of Bench-Capon et al. on argumentation applies a coherence-based mechanism for practical reasoning systems [Dunne and Bench-Capon, 2002]. What is evident from the analysis so far is the lack of a general coherence framework that can accommodate applications in diverse scenarios. We in this book, provide a computational framework that is intended to fill this gap.

## 2.5 Concluding Remarks

This chapter provides the relevant literature and background for this book. Starting from the motivational theories of agency, we have covered some of the significant literature on autonomous agent architectures and their application in autonomous normative agents and normative MAS. We also explored ap-

plications of coherence in linguistic and legal theories. The literature survey highlights the need for cognitive architectures that are more flexible and capable of adapting to situational changes. A specific need to have autonomous conflict resolution mechanisms in normative agents has clearly emerged. Likewise the discussion on the traditional argumentation systems suggest a need for greater amount of flexibility in deliberations and consensus seeking mechanisms.

In analysing the theory of coherence, we have seen that coherence-based reasoning and decision making can potentially fill this gap. We have seen that coherence-based reasoning is significantly different from other traditional approaches in that coherence-maximisation seeks to discover preferences dynamically from constraints that exist among pairs of cognitive elements. The most preferred action chosen in this manner represents the best choice for an agent and reflects changes in preferences over time. Discussing coherence-based reasoning in the presence of conflicting motivations, we see that it is particularly suited for normative reasoning where conflicting motivations due to norms and personal goals are likely.

This discussion ends Part I of the book and now we move on to discuss the technical details of the coherence framework and the coherence-based architecture that forms the core of the dissertation. We also take a specific type of coherence, namely the deductive coherence and analyse its formal properties. This gives a clear method and a computational realisation of coherence which is essential in implementing coherence-driven agents. The coherence-based architecture is then discussed with experimental evaluation of the feasibility of coherence-driven agents.

# Part II

# Framework and Architecture

# Chapter 3

# Coherence Framework and Deductive Coherence

*"Everything should be made as simple as possible,*
*but not one bit simpler."*
Albert Einstein (1879 - 1955)

In Part I of the book, we dwelled upon our motivations for introducing more flexibility and adaptability in autonomous agent architectures. We further justified the choice of the theory of coherence and placed it in the context of other motivational theories and agent architectures. In this chapter, we introduce the formalisation of coherence by defining a coherence framework along with certain computable functions which will form the base of coherence-based agent architecture proposed in later chapters. We also specialise the coherence framework for *deductive coherence graphs* in which the coherence values between pairs of nodes are due to a *deductive coherence function*. A deductive coherence function is defined based on the principles of Thagard. We show that the defined function indeed satisfy Thagard's principles.

## 3.1 Generic Coherence Framework

The framework is based on Thagard's theory of coherence as maximising constraint satisfaction [Thagard, 2002]. As mentioned in the introduction and background, the theory of coherence is based on the underlying assumption that pieces of information can be associated with each other, the association being either positive or negative. Since we are interested in studying these associations, we use a graph structure to model these associations. When pieces of information are cognitive elements, a graph representation enables in making association between cognitive elements explicit, differing substantially from other approaches that extend BDI architecture [Broersen et al., 2002, Pasquier et al., 2006]. With this approach, coherence is treated as a fundamental property that an agent

strives to maintain. Coherence graphs and the computable functions to determine coherence of a graph are introduced in this section.

### 3.1.1 Coherence Graphs

Nodes in a coherence graph represent pieces of information for which we want to estimate coherence. Examples of such pieces of information are propositions representing concepts, actions or mental states both atomic and complex, graded and absolute. Edges between nodes may be associated with a strength, represented by a function $\zeta$, which is derived from an underlying relation between the pieces of information. That is, if two pieces of information are related through an *explanation*, then the function $\zeta$ assigns a positive strength to the edge connecting the two. Based on Thagard's classification of the types of coherence, we have different $\zeta$ functions. Values of function $\zeta$ may be negative or positive. Note that a zero strength on an edge implies that the two pieces of information are unrelated, which is equivalent to not having the edge. Hence, we only consider nonzero strength values on edges. Thagard's principles may be used to define a function $\zeta$ for each of the types of coherence (see Section 3.2).

**Definition 3.1.1** *A* coherence graph *is an edge-weighted undirected graph* $g = \langle V, E, \zeta \rangle$*, where*

1. $V$ *is a finite set of nodes representing pieces of information.*
2. $E \subseteq V^{(2)}$ *(where* $V^{(2)} = \{\{v, w\} \mid v, w \in V\}$*) is a finite set of edges representing the coherence or incoherence between pieces of information.*
3. $\zeta : E \to [-1, 1] \setminus \{0\}$ *is an edge-weighted function that assigns a value to the coherence between pieces of information, and which we shall call a* coherence function

*Let* $\mathcal{G}$ *denote the set of all possible coherence graphs.*

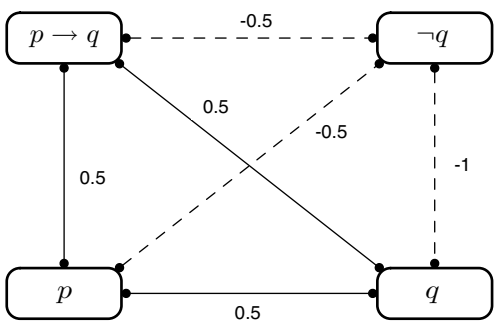


Figure 3.1: An example of a coherence graph

We consider a running example as in Figure 3.1, which will help us to illustrate the concepts as we define them. The graph in the example captures deductive coherence as yielded by classical propositional deduction. As we gradually build our framework, we shall see how these coherence values arise.

### 3.1.2 Calculating Coherence

According to the theory of coherence, if a piece of information is chosen as accepted (or declared true), pieces of information contradicting it are most likely rejected (or declared false) while those supporting it and getting support from it are most likely accepted (or declared true). The important problem is not to find a piece of information that gets accepted, but to know whether more than one piece of information or a set of them can be accepted together. Hence, a coherence problem is to partition the nodes of a coherence graph into two sets (accepted $\mathcal{A}$, and rejected $V \setminus \mathcal{A}$) in such a way as to maximise the satisfaction of constraints. A positive constraint between two nodes is said to be satisfied if both nodes are either in the accepted set or both in the rejected set. Similarly, a negative constraint is satisfied if one of them is in the accepted set while the other is in the rejected set. We express these formally in the following definitions.

**Definition 3.1.2** *Given a coherence graph $g = \langle V, E, \zeta \rangle$, and a partition $(\mathcal{A}, V \setminus \mathcal{A})$ of $V$, the* set of satisfied constraints $C_{\mathcal{A}} \subseteq E$ *is given by*

$$C_{\mathcal{A}} = \left\{ \{v, w\} \in E \;\middle|\; \begin{array}{l} v \in \mathcal{A} \text{ iff } w \in \mathcal{A}, \text{ when } \zeta(\{v, w\}) > 0 \\ v \in \mathcal{A} \text{ iff } w \notin \mathcal{A}, \text{ when } \zeta(\{v, w\}) < 0 \end{array} \right\}$$

*All other constraints (in $E \setminus C_{\mathcal{A}}$) are said to be* unsatisfied.

To illustrate this, consider the partition as in Figure 3.2. We see that $\{p \rightarrow q, \neg q\}$, $\{p \rightarrow q, p\}$ and $\{p, \neg q\}$ are the satisfied constraints. Now we define the coherence-maximising partition as the partition that maximises the satisfaction of constraints. We first define the strength of a partition as the sum over the strengths (the $\zeta$ values) of all the satisfied constraints divided by the number of edges in the graph. Then the coherence-maximising partition is that which maximises the strength, and the coherence value of the graph is defined as the strength of this partition.

**Definition 3.1.3** *Given a coherence graph $g = \langle V, E, \zeta \rangle$,* the strength of a partition $(\mathcal{A}, V \setminus \mathcal{A})$ *of $V$ is given by*

$$\sigma(g, \mathcal{A}) = \frac{\sum_{\{v,w\} \in C_{\mathcal{A}}} \mid \zeta(\{v, w\}) \mid}{\mid E \mid}$$

For the partition in Figure 3.2, the strength is 0.25. Notice that, by Definitions 3.1.2 and 3.1.3,

$$\sigma(g, \mathcal{A}) = \sigma(g, V \setminus \mathcal{A}) \ . \tag{3.1}$$

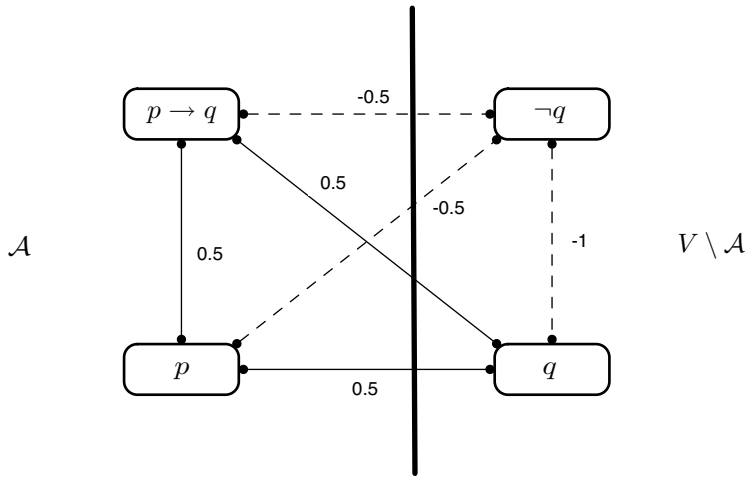


Figure 3.2: A partition of a coherence graph

**Definition 3.1.4** *Given a coherence graph $g = \langle V, E, \zeta \rangle$,* the coherence of $g$ *is given by*

$$\kappa(g) = \max_{\mathcal{A} \subseteq V} \sigma(g, \mathcal{A})$$

*If for some partition $(\mathcal{A}, V \setminus \mathcal{A})$ of $V$, the strength of the partition is maximal (i.e., $\kappa(g) = \sigma(g, \mathcal{A})$ then the set $\mathcal{A}$ is called the* accepted *set and $V \setminus \mathcal{A}$ the* rejected *set of the partition.*

Due to Equation 3.1, the accepted set $\mathcal{A}$ is never unique for a coherence graph. Moreover, there could be other partitions that generate the same value for $\kappa(g)$. Here we mention some possible criteria for selecting an accepted set among the alternatives. If $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ are sets from all those partitions that maximise coherence of the graph $g$, first we may choose the accepted set to which the intuitively obvious propositions belong. This is based on one of Thagard's principles (which we will formalise in the next definition) on deductive coherence [Thagard, 2002], namely that *intuitively obvious propositions have an acceptability on their own.* And lastly, an accepted set with more number of elements could be preferred to another with less.

The coherence maximising partition for the example is as in Figure 3.3. With this partition we see that maximum constraints are satisfied, and the coherence of the graph (the maximum strength of the partition in Figure 3.3.) is $0.58\overline{3}$.

For a set of pieces of information, we are now equipped to find coherence maximising partitions and in most cases a unique accepted set of pieces of information with the help of the functions introduced above. As mentioned earlier, with coherence maximisation, agent reasoning and decision making is formulated as a classification problem. A simplistic scenario is when an agent has two

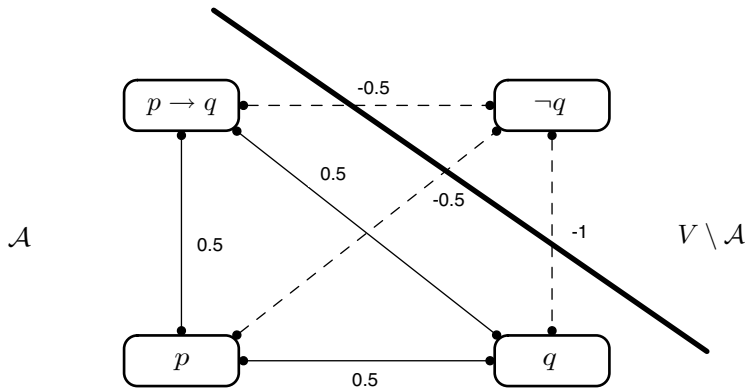


Figure 3.3: Coherence-maximising partition of a coherence graph

alternative to chose from, and if only one of them appear in the accepted set, then the agent choses the alternative that is in the accepted set.

In the next section, we specialise the coherence framework for deductive coherence graphs and introduce methods to construct such coherence graphs given a set of pieces of information. This specialisation makes a coherence framework fully computational and differentiates this proposal from the state of the art.

## 3.2 Deductive Coherence

So far we have introduced some of the general computable functions of the coherence framework under the assumption that a coherence graph already exists. For this framework to be fully computational, it is necessary to define how a coherence graph can be constructed. Given a set of pieces of information (and possibly some associated confidence degrees), we need to define a coherence function $\zeta$ that assigns a coherence value to pairs of pieces of information. As the nature of relationship between two pieces of information can vary greatly, we may need to define more than one coherence function. But for each type of coherence, only one such relationship is evaluated. That is, for explanatory coherence for instance, two pieces of information are coherent only if they are related by an explanation.

In this section we introduce deductive coherence, and define a *deductive coherence graph* whose nodes are propositions and pairs of nodes are related by a *deductive coherence function* $\zeta$ yielded by propositional logical deduction. We choose deductive coherence among the different types of coherence because logical deduction has a sound theoretical basis and has well defined rules in order to start with a formalisation of coherence. We first derive a deductive coherence function in adherence with Thagard's principles and in Chapter 4 analyse this function in the context of structural and internal connectives. The latter helps

us to further derive coherence values between those propositions that are not directly related by deduction.

**Definition 3.2.1** *A* deductive coherence graph *is an edge-weighted undirected graph* $g = \langle V, E, \zeta \rangle$*, where*

1. $V$ *is a finite set of nodes representing pieces of information expressed as propositions.*
2. $E \subseteq V^{(2)}$ *(where* $V^{(2)} = \{\{v, w\} \mid v, w \in V\}$*) is a finite set of edges representing the coherence or incoherence between the propositions.*
3. $\zeta : E \to [-1, 1] \setminus \{0\}$ *is a deductive coherence function that assigns a value to the coherence between pairs of propositions.*

### 3.2.1 Deductive Coherence Function

Here we discuss Thagar'd principles and define a deductive coherence function adhering to these principles. Thagard introduces the notion of deductive coherence by means of a set of principles [Thagard, 2002]:

| Number | Principle |
|---|---|
| 1. | Deductive coherence is a symmetric relation. |
| 2. | A proposition coheres with propositions that are deducible from it. |
| 3. | Propositions that together are used to deduce some other proposition cohere with each other. |
| 4. | The more hypotheses it takes to deduce something, the less the degree of coherence. |
| 5. | Contradictory propositions are incoherent with each other. |
| 6. | Propositions that are intuitively obvious have a degree of acceptability on their own. |
| 7. | The acceptability of a proposition in a system of propositions depends on its coherence with them. |

Table 3.1: Thagard's Principles on Deductive Coherence

Before going into the details of Thagard's principles, it is important to note that these principles were proposed taking into account a context or —in logical terminology— a theory $\mathcal{T}$. Examples of such theories may be the theory of arithmetic while proving theorems in mathematics, or legal laws while making legal judgements. In the context of autonomous normative agents, the set of rules and observations about the context is this theory. To be rigorous we should call

$\mathcal{T}$ a finite theory presentation. However, to avoid lengthy phrases, we will often call it just a theory. Assuming bounded rationality for our agents, $\mathcal{T}$ is not closed under deduction. We essentially see the process of coherence maximisation as a process of theory revision. That is, each time the agent encounters a new piece of information $\beta$ (a new norm, a new belief, etc.), it tries to relate it to the theory presentation it has. The new information can influence $\mathcal{T}$ in the following ways:

1. *Extend $\mathcal{T}$: $\beta$ helps to deduce propositions that were not deducible before.*

2. *Extend $\mathcal{T}$: $\beta$ is deducible from $\mathcal{T}$.*

3. *Modify $\mathcal{T}$: $\beta$ is in a deduction relation with some propositions in $\mathcal{T}$, however, contradicts some other.*

The coherence function we propose here is in the context of a theory $\mathcal{T}$ and is motivated to aid this process of theory revision. That is, by computing the deductive coherence values between pairs of elements of the theory and the new piece of information $\beta$ we essentially have a new coherence graph. The coherence of the graph then modifies the theory by partitioning the theory elements into accepted and rejected sets. This said however, it is hard to compare coherence as a theory revision method with other theory revision methods that follows the AGM postulates [Koons, 2009]. This is because, coherence based theory revision does not follow the AGM postulates for the simple reason that using coherence as a theory revision mechanism, we are accepting to have inconsistencies in the theory. Coherence maximisation only *maximises* satisfaction of constraints (minimise inconsistencies) but not eliminates it.

Since Thagard's principles on deductive coherence are based on an intuitive notion of deduction, we try to make it generic by basing our coherence function on multiset deduction relations (MDR). The concept of a multiset is a generalisation of the concept of a set. Intuitively speaking, we can regard a multiset as a set in which the number of times each element occurs is significant, but not the order of the elements. The introduction of multisets in our framework will allow us to deal more adequately with logics such as linear logics, relevance logics or multi-valued logics[1]. We assume that all MDRs we deal with are finitary and decidable. These MDRs are often called *simple consequence relations* [Avron, 1991]. We define an MDR as follows:

**Definition 3.2.2** *Given a logical language $L$, a MDR on $L$, is a binary relation $\vdash$ between finite multisets of formulas of $L$ such that, for all $\Gamma, \Gamma_1, \Gamma_2, \Sigma_1, \Sigma_2 \subseteq L$ and for all $\gamma \in L$:*

**Reflexivity:** $\Gamma \vdash \Gamma$

**Transitivity:** *If* $\Gamma_1 \vdash \Sigma_1, \gamma$ *and* $\gamma, \Gamma_2 \vdash \Sigma_2$*, then* $\Gamma_1, \Gamma_2 \vdash \Sigma_1, \Sigma_2$

[1]These logics are more relevant while dealing with deductive coherence. It will become apparent when we discuss Thagard's principles.

We denote by $\vdash \beta$ the fact that $\beta$ can be deduced from the empty multiset, and we denote by $\Gamma \vdash$ the fact that the multiset $\Gamma$ has as consequence the empty multiset. For example, in case that $L$ is classical propositional logic, $\vdash \beta$ means that $\beta$ is a tautology and $\Gamma \vdash$ means that the multiset $\Gamma$ is inconsistent.

We use Thagard's principles to relate an MDR with a coherence function $\zeta$. Below we analyse each of the principles and discuss the way we incorporate the principle in our definition of $\zeta$.

1. To make sure that the deductive coherence function is symmetric, we first define two support functions between pairs of propositions $\alpha$ and $\beta$, which evaluate the support of $\alpha$ in deriving $\beta$ and vice versa. Thus the support function is assymetric. Later the maximum of the two support function values is defined as the value of the coherence function which makes it a symmetric function. The maximum is chosen due to the fact that, even if there is only a deduction relation in one of the directions, there is a deductive coherence between the two propositions.

2. Thagard uses an intuitive notion of deduction for principle 2. If this principle were used with a logic that has weakening, then the result would be that every proposition coheres with every other. The logics like relevant logics which has no weakening is safe from these undesired conclusions and may be more suited to model deductive coherence. It is for this reason that we use multiset deductions which is inclusive of logics like relevant, linear and many valued. Further, in general to make sure that only relevant deductions are made to relate through coherence, we take advantage of the context of a theory. Hence we make it the case that only those deductions that use the theory are valid for considering for establishing a coherence relation.

3. Principles 2 and 3 capture the fact that there are certain positive coherence relations between premises, and between each of the premises and the conclusion. Since we want to focus only on those deductions that fall in the context of the theory, we shall define the coherence only between those formulas that are either in the theory or are subformulas thereof (hence Definition 3.2.3 below).

4. Principle 4 gives an indication of the magnitude of coherence. It states that the magnitude of coherence decreases with the increasing number of premises required. Hence we make the coherence value inversely proportional to the number of premises.

5. Principle 5 discusses the case of contradiction. We model contradiction similar to deduction by extending Principle 4 to guide the magnitude of contradiction.

6. Principle 6 is not directly captured in the definition of $\zeta$. That is, other than the fact that intuitively obvious propositions may be involved in more

deductive relations with the rest of the theory elements, we do not incorporate this in the definition. However, to disambiguate among many possible accepted sets, we give a priorities for candidate sets on the basis of the presence of intuitively obvious propositions.

7. And finally, Principle 7 stresses the basic notion of coherence, namely that if anything is accepted, it is because accepting it improves the coherence of the system. Therefore, the theory $\mathcal{T}$ is also part of our coherence graph, and its acceptance is only with respect to coherence maximisation.

We first formalise Thagard's principles for classical propositional logic. Principles 2–7 are formalised in terms of a *support function* $\eta$ with respect to a finite theory presentation $\mathcal{T}$, and then we use this function to define the coherence function $\zeta$ in a way that captures the symmetry of coherence (Principle 1). Later, in Chapter 5, we generalise these functions for a many-valued logic, reinterpreting Thagard's principles appropriately.

**Definition 3.2.3** *Let $L$ be a logical language and let $\mathcal{T} \subseteq L$ be a finite theory presentation. We call $\mathcal{T}^\bullet$ the* closure of $\mathcal{T}$ under subformulas *when $\alpha' \in \mathcal{T}^\bullet$ if and only if there is an $\alpha \in \mathcal{T}$ such that $\alpha'$ is a subformula of $\alpha$.*

**Definition 3.2.4** *Let $L$ be a logical language and $\vdash$ be an MDR for $L$. Let $\mathcal{T} \subseteq L$ be a finite theory presentation. A* support function $\eta_{\mathcal{T}} : \mathcal{T}^\bullet \times \mathcal{T}^\bullet \to [-1, 1] \setminus \{0\}$ *with respect to $\mathcal{T}$ is given by:*

$$\eta_{\mathcal{T}}(\alpha, \beta) = \begin{cases} \max \begin{cases} \frac{1}{|\Gamma|+1} & \text{where } \Gamma \text{ is the smallest subset of } \mathcal{T}^\bullet \text{ such that} \\ & \Gamma, \alpha \vdash \beta \text{ and } \Gamma \nvdash \beta \text{ and } \alpha \nvdash \\ \frac{1}{|\Gamma|+2} & \text{where } \Gamma \text{ is the smallest subset of } \mathcal{T}^\bullet \text{ such that} \\ & \exists \gamma \in \mathcal{T}^\bullet \text{ such that } \Gamma, \alpha, \beta \vdash \gamma \text{ and} \\ & \Gamma, \alpha \nvdash \gamma \text{ and } \Gamma, \beta \nvdash \gamma \text{ and } \gamma \nvdash \\ \frac{-1}{|\Gamma|+1} & \text{where } \Gamma \text{ is the smallest subset of } \mathcal{T}^\bullet \text{ such that} \\ & \Gamma, \alpha, \beta \vdash \text{ and } \Gamma, \alpha \nvdash \text{ and } \Gamma, \beta \nvdash \end{cases} \\ \text{undefined, otherwise} \end{cases}$$

Since deductive coherence is symmetric, we now set the value of the deductive coherence between two propositions to be the greatest value of the support function for these propositions. Due to this, even if there may only be a deduction relation in one direction, there will be deductive coherence in both directions. Note that both the support function and the deductive coherence function are partial functions. This is because we interpret zero coherence as the propositions not being related.

**Definition 3.2.5** *Let $L$ be a logical language and $\vdash$ be an MDR for $L$. Let $\mathcal{T} \subseteq L$ be a finite theory presentation. Let $\eta_{\mathcal{T}} : \mathcal{T}^\bullet \times \mathcal{T}^\bullet \to [-1, 1] \setminus \{0\}$*

*be a support function with respect to $\mathcal{T}$. A deductive coherence function $\zeta_{\mathcal{T}} : (\mathcal{T}^{\bullet})^{(2)} \to [-1,1] \setminus \{0\}$ with respect to $\mathcal{T}$ is given by:*

$$\zeta_{\mathcal{T}}(\{\alpha,\beta\}) = \begin{cases} \max\{\eta_{\mathcal{T}}(\alpha,\beta), \eta_{\mathcal{T}}(\beta,\alpha)\} & \text{if } \eta_{\mathcal{T}}(\alpha,\beta) \text{ and } \eta_{\mathcal{T}}(\beta,\alpha) \text{ are defined} \\ \eta_{\mathcal{T}}(\alpha,\beta) & \text{if } \eta_{\mathcal{T}}(\alpha,\beta) \text{ is defined and } \eta_{\mathcal{T}}(\beta,\alpha) \text{ is undefined} \\ \text{undefined} & \text{if } \eta_{\mathcal{T}}(\alpha,\beta) \text{ and } \eta_{\mathcal{T}}(\beta,\alpha) \text{ are undefined} \end{cases}$$

*Our example of Figure 3.1 assumes that $\mathcal{T} = \{p \to q, \neg q\}$, and consequently $\mathcal{T}^{\bullet} = \{p \to q, \neg q, p, q\}$. The only relevant deductions using formulas of $\mathcal{T}^{\bullet}$ (and assuming classical propositional deduction) are:*

$$\begin{aligned} p \to q, p &\vdash q \\ q, \neg q &\vdash \\ p \to q, p, \neg q &\vdash \end{aligned}$$

*Therefore, we have that*

$$\begin{aligned} \eta_{\mathcal{T}}(p,q) &= \frac{1}{|\{p \to q\}| + 1} = 0.5 \\ \eta_{\mathcal{T}}(p \to q, q) &= \frac{1}{|\{p\}| + 1} = 0.5 \\ \eta_{\mathcal{T}}(p \to q, p) = \eta_{\mathcal{T}}(p, p \to q) &= \frac{1}{|\emptyset| + 2} = 0.5 \\ \eta_{\mathcal{T}}(q, \neg q) = \eta_{\mathcal{T}}(\neg q, q) &= \frac{-1}{|\emptyset| + 1} = -1 \\ \eta_{\mathcal{T}}(p, \neg q) = \eta_{\mathcal{T}}(\neg q, p) &= \frac{-1}{|\{p \to q\}| + 1} = -0.5 \\ \eta_{\mathcal{T}}(p \to q, \neg q) = \eta_{\mathcal{T}}(\neg q, p \to q) &= \frac{-1}{|\{p\}| + 1} = -0.5 \end{aligned}$$

*and consequently,*

$$\begin{aligned} \zeta_{\mathcal{T}}(\{p,q\}) &= 0.5 \\ \zeta_{\mathcal{T}}(\{p \to q, q\}) &= 0.5 \\ \zeta_{\mathcal{T}}(\{p \to q, p\}) &= 0.5 \\ \zeta_{\mathcal{T}}(\{q, \neg q\}) &= -1 \\ \zeta_{\mathcal{T}}(\{p, \neg q\}) &= -0.5 \\ \zeta_{\mathcal{T}}(\{p \to q, \neg q\}) &= -0.5 \end{aligned}$$

*For all remaining pairs of formulas from $\mathcal{T}^{\bullet}$, the value of $\zeta_{\mathcal{T}}$ is undefined.*

**Proposition 3.2.1** *The deductive coherence function $\zeta_{\mathcal{T}}$ as defined in Definition 3.2.5 satisfies Thagard's principles of deductive coherence (see Section 3.2.1).*

**Proof 3.2.2**
*For all $\zeta_{\mathcal{T}}$, coherence is symmetric by construction, which satisfies Principle 1.*

*Let $\Gamma \subseteq \mathcal{T}^{\bullet}$ and $\alpha, \beta \in \mathcal{T}^{\bullet}$ such that $\Gamma, \alpha \vdash \beta$ and $\Gamma \nvdash \beta$ and $\alpha \nvdash$. Then $\eta_{\mathcal{T}}(\alpha, \beta) > 0$, and consequently, $\zeta_{\mathcal{T}}(\{\alpha, \beta\}) > 0$, which satisfies Principle 2.*

*Let $\Gamma \subseteq \mathcal{T}^{\bullet}$ and $\alpha, \beta, \gamma \in \mathcal{T}^{\bullet}$ such that $\Gamma, \alpha, \beta \vdash \gamma$ and $\Gamma, \alpha \nvdash \gamma$ and $\Gamma, \beta \nvdash \gamma$ and $\gamma \nvdash$. Then $\eta_{\mathcal{T}}(\alpha, \beta) > 0$, and consequently, $\zeta_{\mathcal{T}}(\{\alpha, \beta\}) > 0$, which satisfies Principle 3.*

*Let $\Gamma_1, \Gamma_2 \subseteq \mathcal{T}^{\bullet}$ with $|\Gamma_1| < |\Gamma_2|$, and let $\alpha_1, \alpha_2, \beta \in \mathcal{T}^{\bullet}$ such that $\Gamma_1, \alpha_1, \vdash \beta$ and there does not exists $\Gamma_1'$ such that $|\Gamma_1'| < |\Gamma_1|$ and $\Gamma_1', \alpha_1, \vdash \beta$, and $\Gamma_2, \alpha_2 \vdash \beta$ and there does not exists $\Gamma_2'$ such that $|\Gamma_2'| < |\Gamma_2|$ and $\Gamma_2', \alpha_2, \vdash \beta$, and $\Gamma_1 \nvdash \beta$ and $\Gamma_2 \nvdash \beta$ and $\alpha_1 \nvdash$ and $\alpha_2 \nvdash$. Then $\eta_{\mathcal{T}}(\alpha_1, \beta) > \eta_{\mathcal{T}}(\alpha_2, \beta)$, and consequently, $\zeta_{\mathcal{T}}(\{\alpha_1, \beta\}) > \zeta_{\mathcal{T}}(\{\alpha_2, \beta\})$, which satisfies Principle 4.*

*Let $\alpha, \beta \in \mathcal{T}^{\bullet}$ such that $\alpha, \beta \vdash$ and $\alpha \nvdash$ and $\beta \nvdash$. Consequently $\eta_{\mathcal{T}}(\alpha, \beta) < 0$, and consequently $\zeta_{\mathcal{T}}(\{\alpha, \beta\}) < 0$, which satisfies Principle 5.*

*Axioms that are intuitively obvious are supposed to be those that belong to the theory $\mathcal{T}$. Let $\Gamma \subseteq \mathcal{T}^{\bullet}$ and $\alpha, \beta \in \mathcal{T}^{\bullet}$ such that $\Gamma, \alpha \vdash \beta$ and $\Gamma \nvdash \beta$ and $\alpha \nvdash$. Then, for all $\gamma \in \Gamma$, $\eta_{\mathcal{T}}(\gamma, \alpha) > 0$. Hence, axioms in $\mathcal{T}$ and its subformulas that participate with other formulas in deduction relations cohere positively with them, having thus a higher degree of acceptability, which satisfies Principle 6.*

*Finally, Definition 7.2 satisfies Principle 7.*

## 3.3 Discussion

In this chapter, we have defined the coherence framework and specialized it for deductive coherence graphs. The coherence framework along with this specialisation enable us to construct deductive coherence graphs and compute coherence maximising partitions of such graphs. This makes our construction fully computational. Further, by proving that the deductive coherence function satisfies all the principles of Thagard, we also show that it is sound. The next chapter is a continuation of current chapter where we explore the logical properties of the deductive coherence function. These properties enable us to compute coherence values between propositions that are otherwise unconnected.

# Chapter 4

# Formalising Coherence: A Proof-Theoretical Approach

The deductive coherence function defined in Chapter 3 enable us to construct deductive coherence graphs given a set of propositional formulas. The function determines coherence values between pairs of propositional formulas that are directly related through deduction. In this chapter, we explore some of the properties of deductive coherence function $\zeta_{\mathcal{T}}$ to determine the values for pairs of formulas related through some of the structural rules and connectives of the underlying logic. Avron in his seminal paper on "Simple Consequence Relations" classifies consequence relations according to their basic connectives [Avron, 1991]. The aim was to establish a syntactic characterisation of consequence relations in terms of the connectives definable in them. With a similar aim of giving a syntactic characterisation of the deductive coherence function, we identify the properties of the support function $\eta_{\mathcal{T}}$ using the properties of the connectives and structural rules. Although these properties are not essential for the understanding of the coherence-based agent architecture introduced in Chapter 5, this analysis helps us to stress the generality of our approach.

## 4.1 Properties of Deductive Coherence Based on MDRs

We can classify logics according to structural rules (such as weakening or monotonicity) and connectives available in it. There are two types of connectives: the *internal* connectives, which transform a given sequent into an equivalent one that has a special required form, and the *combining* connectives, which combine two sequents into one. For instance, classical propositional logic is monotonic, satisfies weakening, has all internal and combining connectives, and makes no difference between them. On the other hand, propositional linear logic is monotonic, has all connectives, but distinguishes between internal and com-

bining ones. Intuitionistic logic differs from classical propositional logic in its implication connective and does not have internal negation.

By Definition 3.2.4, the function $\eta_{\mathcal{T}}$ is defined for formulas of $\mathcal{T}^{\bullet} \subseteq L$ related through an MDR in the form $\Gamma, \alpha \vdash \beta$. Hence, we express the deduction relation in this single-conclusioned form so that we can find properties of function $\eta_{\mathcal{T}}$ between different formulas of the premises and conclusion, using the properties of the connectives.[1]

### 4.1.1 Combining Conjunction

A conjunction $\wedge$ is *combining* iff, for all $\Gamma, \Sigma \subseteq L$ and $\alpha, \beta \in L$,

$$\Gamma \vdash \Sigma, \alpha \wedge \beta \quad \textit{iff} \quad \Gamma \vdash \Sigma, \alpha \textit{ and } \Gamma \vdash \Sigma, \beta$$

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^{\bullet}$ and $\alpha \wedge \beta, \gamma \in \mathcal{T}^{\bullet}$ such that $\Gamma \nvdash \alpha$ and $\Gamma \nvdash \beta$ and $\Gamma \nvdash \alpha \wedge \beta$ and $\alpha \wedge \beta \nvdash$ and $\gamma \nvdash$.

1. If $\Gamma, \gamma \vdash \alpha \wedge \beta$ then $\eta(\gamma, \alpha \wedge \beta) > 0$ and, since $\Gamma, \gamma \vdash \alpha$ and $\Gamma, \gamma \vdash \beta$, we have that $\eta(\gamma, \alpha) \geq \eta(\gamma, \alpha \wedge \beta)$ and $\eta(\gamma, \beta) \geq \eta(\gamma, \alpha \wedge \beta)$.

2. If $\Gamma, \gamma \vdash \alpha$ and $\Gamma, \gamma \vdash \beta$ then $\eta(\gamma, \alpha) > 0$ and $\eta(\gamma, \beta) > 0$. Let their values be $\frac{1}{n}$ and $\frac{1}{m}$, respectively. Since $\Gamma, \gamma \vdash \alpha \wedge \beta$, we further have that $\eta(\gamma, \alpha \wedge \beta) \geq \frac{1}{n+m-1}$.

3. Finally, $\eta(\alpha \wedge \beta, \alpha) = 1$ and $\eta(\alpha \wedge \beta, \beta) = 1$.

### 4.1.2 Internal Conjunction

A conjunction $\circ$ is *internal* iff, for all $\Gamma, \Sigma \subseteq L$ and $\alpha, \beta \in L$,

$$\Gamma, \alpha, \beta \vdash \Sigma \quad \textit{iff} \quad \Gamma, \alpha \circ \beta \vdash \Sigma$$

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^{\bullet}$ and $\alpha \circ \beta, \sigma \in \mathcal{T}^{\bullet}$ such that $\Gamma \nvdash \sigma$ and $\Gamma, \alpha \nvdash \sigma$ and $\Gamma, \beta \nvdash \sigma$ and $\alpha \circ \beta \nvdash$ and $\alpha \nvdash$ and $\beta \nvdash$.

1. If $\Gamma, \alpha \circ \beta \vdash \sigma$ then $\eta(\alpha \circ \beta, \sigma) > 0$. Let its value be $\frac{1}{n}$. Since $\Gamma, \alpha, \beta \vdash \sigma$, we have that $\eta(\alpha, \sigma) \geq \frac{1}{n+1}$ and $\eta(\alpha, \sigma) \geq \frac{1}{n+1}$.

2. If $\Gamma, \alpha, \beta \vdash \sigma$ then $\eta(\alpha, \sigma) > 0$ and $\eta(\beta, \sigma) > 0$. Let their values be $\frac{1}{n}$ and $\frac{1}{m}$, respectively. Since $\Gamma, \alpha \circ \beta \vdash \sigma$, we further have that $\eta(\alpha \circ \beta, \sigma) \geq \frac{1}{n+m-3}$.

3. Finally, $\eta(\alpha, \alpha \circ \beta) \geq \frac{1}{2}$ and $\eta(\beta, \alpha \circ \beta) \geq \frac{1}{2}$ and $\eta(\alpha, \beta) \geq \frac{1}{2}$.

[1] For convenience, in the rest of this subsection we shall drop the subindex of $\eta_{\mathcal{T}}$; however, it should be noted that it is always evaluated with respect to a finite theory presentation $\mathcal{T}$.

### 4.1.3 Combining Disjunction

A disjunction $\vee$ is *combining* iff, for all $\Gamma, \Sigma \subseteq L$ and $\alpha, \beta \in L$,

$$\Gamma, \alpha \vee \beta \vdash \Sigma \quad \textit{iff} \quad \Gamma, \alpha \vdash \Sigma \textit{ and } \Gamma, \beta \vdash \Sigma$$

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^\bullet$ and $\alpha \vee \beta, \sigma \in \mathcal{T}^\bullet$ such that $\Gamma \nvdash \sigma$ and $\alpha \vee \beta \nvdash$ and $\alpha \nvdash$ and $\beta \nvdash$.

1. If $\Gamma, \alpha \vee \beta \vdash \sigma$ then $\eta(\alpha \vee \beta, \sigma) > 0$ and, since $\Gamma, \alpha \vdash \sigma$ and $\Gamma, \beta \vdash \sigma$, we have that $\eta(\alpha, \sigma) \geq \eta(\alpha \vee \beta, \sigma)$ and $\eta(\beta, \sigma) \geq \eta(\alpha \vee \beta, \sigma)$.
2. If $\Gamma, \alpha \vdash \sigma$ and $\Gamma, \beta \vdash \sigma$ then $\eta(\alpha, \sigma) > 0$ and $\eta(\beta, \sigma) > 0$. Let their values be $\frac{1}{n}$ and $\frac{1}{m}$, respectively. Since $\Gamma, \alpha \vee \beta \vdash \sigma$, we further have that $\eta(\alpha \vee \beta, \sigma) \geq \frac{1}{n+m-1}$.
3. For all $\gamma \in \mathcal{T}^\bullet$, we have that $\eta(\gamma, \alpha \vee \beta) = -1$ iff both $\eta(\gamma, \alpha) = -1$ and $\eta(\gamma, \beta) = -1$.
4. Finally, $\eta(\alpha, \alpha \vee \beta) = 1$ and $\eta(\beta, \alpha \vee \beta) = 1$.

### 4.1.4 Internal Disjunction

A disjunction $+$ is *internal* iff, for all $\Gamma, \Sigma \subseteq L$ and $\alpha, \beta \in L$,

$$\Gamma \vdash \Sigma, \alpha, \beta \quad \textit{iff} \quad \Gamma \vdash \Sigma, \alpha + \beta$$

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^\bullet$ and $\alpha + \beta, \gamma \in \mathcal{T}^\bullet$ such that $\Gamma \nvdash \alpha$ and $\Gamma \nvdash \alpha + \beta$ and $\alpha \nvdash$ and $\beta \nvdash$ and $\gamma \nvdash$. Further, let $\vdash$ satisfy Weakening.[2]

1. We distinguish three cases:
   - If $\eta(\gamma, \alpha) > 0$ because $\Gamma, \gamma \vdash \alpha$, and $\eta(\gamma, \beta)$ is undefined, then, since $\Gamma, \gamma \vdash \alpha, \beta$ and hence $\Gamma, \gamma \vdash \alpha + \beta$, we have that $\eta(\gamma, \alpha + \beta) \geq \eta(\gamma, \alpha)$;
   - If $\eta(\gamma, \alpha)$ is undefined, and $\eta(\gamma, \beta) > 0$ because $\Gamma, \gamma \vdash \beta$, then, since $\Gamma, \gamma \vdash \alpha, \beta$ and hence $\Gamma, \gamma \vdash \alpha + \beta$, we have that $\eta(\gamma, \alpha + \beta) \geq \eta(\gamma, \beta)$;
   - If both $\eta(\gamma, \alpha) > 0$ and $\eta(\gamma, \beta) > 0$ because both $\Gamma, \gamma \vdash \alpha$ and $\Gamma, \gamma \vdash \beta$, then, since $\Gamma, \gamma \vdash \alpha, \beta$ and hence $\Gamma, \gamma \vdash \alpha + \beta$, we have that $\eta(\gamma, \alpha + \beta) \geq \max\{\eta(\gamma, \alpha), \eta(\gamma, \beta)\}$.
2. Finally, $\eta(\alpha, \alpha + \beta) = 1$ and $(\beta, \alpha + \beta) = 1$.

### 4.1.5 Combining Implication

An implication $\supset$ is *combining* iff, for all $\Gamma, \Sigma \subseteq L$ and $\alpha, \beta \in L$,

$$\Gamma, \alpha \supset \beta \vdash \Sigma \quad \textit{iff} \quad \Gamma \vdash \Sigma, \alpha \textit{ and } \Gamma, \beta \vdash \Sigma$$

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^\bullet$ and $\alpha \supset \beta, \sigma \in \mathcal{T}^\bullet$ such that $\Gamma \nvdash \sigma$ and $\alpha \supset \beta \nvdash$ and $\beta \nvdash$.

[2] An MDR $\vdash$ satisfies **Weakening** if, for all $\Gamma, \Gamma', \Sigma, \Sigma' \subseteq L$, if $\Gamma \vdash \Sigma$ then $\Gamma, \Gamma' \vdash \Sigma, \Sigma'$.

1. If $\Gamma, \alpha \supset \beta \vdash \sigma$ then $\eta(\alpha \supset \beta, \sigma) > 0$ and, since $\Gamma, \beta \vdash \sigma$, we have that $\eta(\beta, \sigma) \geq \eta(\alpha \supset \beta, \sigma)$.

2. Finally, $\eta(\beta, \alpha \supset \beta) = 1$.

### 4.1.6 Internal Implication

An implication $\rightarrow$ is *internal* iff, for all $\Gamma, \Sigma \subseteq L$ and $\alpha, \beta \in L$,

$$\Gamma, \alpha \vdash \Sigma, \beta \quad \textit{iff} \quad \Gamma \vdash \Sigma, \alpha \rightarrow \beta$$

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^\bullet$ and $\alpha \rightarrow \beta, \gamma \in \mathcal{T}^\bullet$ such that $\Gamma, \gamma \nvdash \beta$ and $\Gamma \nvdash \alpha \rightarrow \beta$ and $\alpha \nvdash$ and $\gamma \nvdash$.

1. If $\Gamma, \gamma \vdash \alpha \rightarrow \beta$ then $\eta(\gamma, \alpha \rightarrow \beta) > 0$. Let its value be $\frac{1}{n}$. Since $\Gamma, \gamma, \alpha \vdash \beta$, we have that $\eta(\gamma, \alpha) \geq \frac{1}{n}$ and $\eta(\gamma, \beta) \geq \frac{1}{n+1}$.

2. If $\Gamma, \gamma, \alpha \vdash \beta$ then $\eta(\gamma, \beta) > 0$. Let its value be $\frac{1}{n}$. Since $\Gamma, \gamma \vdash \alpha \rightarrow \beta$, we have that $\eta(\gamma, \alpha \rightarrow \beta) \geq \frac{1}{n-1}$.

3. If $\alpha \rightarrow \beta \in \mathcal{T}$ then $\eta(\alpha, \beta) \geq \frac{1}{2}$.

4. Finally, $\eta(\alpha \rightarrow \beta, \beta) \geq \frac{1}{2}$.

### 4.1.7 Internal Negation

A negation $\neg$ is *internal* iff, for all $\Gamma, \Sigma \subseteq L$ and $\alpha \in L$,

$$\Gamma, \alpha \vdash \Sigma \quad \textit{iff} \quad \Gamma \vdash \Sigma, \neg\alpha$$

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^\bullet$ and $\neg\alpha, \gamma \in \mathcal{T}^\bullet$ such that $\Gamma, \alpha \nvdash$ and $\Gamma, \gamma \nvdash$.

1. if $\eta(\gamma, \alpha) < 0$ then $\Gamma, \gamma, \alpha \vdash$, and consequently $\Gamma, \gamma \vdash \neg\alpha$ and hence $\eta(\gamma, \neg\alpha) \geq -\eta(\gamma, \alpha)$.

2. If $\Gamma, \gamma \vdash \neg\alpha$ then $\eta(\gamma, \neg\alpha) > 0$, and since $\Gamma, \gamma, \alpha \vdash$ we have that $\eta(\gamma, \alpha) \leq -\eta(\gamma, \neg\alpha)$.

3. $\eta(\neg\alpha, \alpha) = -1$

## 4.2 Concluding Remarks

Analysing coherence proof-theoretically clarifies properties of coherence as a logical relation and affirms the use of logic in interpreting coherence. It further clarifies some of the properties of logic that aligns with the semantic interpretation of coherence. For example, a logic with weakening rule may not be a good candidate for modelling coherence since with weakening, it is easy to show

that every formula is coherently related to every other formula. Hence, paraconsistent logics such as relevant logic are mode ideal to model coherence when compared to classical logic [Paoli, 2002].

In the following chapter, we define a coherence-based architecture based on the coherence framework. Since we are interested in introducing as much flexibility, we will further generalise the deductive coherence functions defined in Chapter 3 to include many-valued logics such as Łukasiewicz logic, which we use to reason about beliefs, desires and intentions of an agent.

# Chapter 5

# Coherence-driven Agents

In this chapter we describe a coherence-based architecture based on the coherence framework developed in Chapters 3 and 4. Later, we define *Coherence-driven agents* to be agents with a coherence-based architecture. We combine a number of formalisms to get a flexible yet expressive architecture, at the cost of apparent complexity. To help readers, we therefore outline the organisation of the chapter and briefly mention the formalisms used in the architecture.

## 5.1 Organisation of the Chapter

Theory of coherence when applied to reasoning, pieces of information may be mapped to cognition (mental attitudes of information, motivation and deliberation as explained in Chapter 2). A coherence maximisation then partitions these cognitive elements into accepted and rejected sets. Since, an agent reasoning is influenced by the way cognitive elements are represented, we would like to have an architecture that has an expressive representation of the same. As is evident from Chapter 2, a BDI architecture has an expressive representation of cognitive elements corresponding to the information, motivational and deliberative attitudes and has successful implementations of the same. Hence, we choose BDI as a base architecture in which we incorporate elements of the coherence framework. To ensure further flexibility, we take the view that agents have an imprecise model of the environment they are in. Hence, we chose the g-BDI architecture (see Section 2.2) which shares this assumption and has a graded representation of the cognitive elements to incorporate uncertainty associated with an agent's world model.

Usually a logic is used such as different flavours of modal logic to model cognition of an agent [Garson, 2009, Blackburn et al., 2006, Rao and Georgeff, 1995]. One of the main motivations for a coherence-based architecture is to make explicit associations among cognitive elements. Merging logics to do is known to be a hard task, and the usual alternative taken to reason between cognition is using multi-context systems(MCS) originally proposed by

Giunchiglia [Giunchiglia and Giunchiglia, 1993, Giunchiglia and Serafini, 1994]. Since the g-BDI architecture already incorporates MCS, we simply adapt the architecture to incorporate the notion of coherence. In this adaptation, contexts have a logic and theories in the logic have a structure in the form of coherence graphs. We need the below notions before we get to define coherence-based agent architecture.

1. Graded model of belief, desire, and intention (g-BDI architecture).
2. MCS architecture with each context consisting of a
   context logic—language, axioms and deduction rules,
   context graph.
3. Definition of a reasoning mechanism across contexts.

In Section 5.2, we detail our adaptation of MCS architecture specifying each of its contexts and the mechanism to reason across contexts. Section 5.3 specify coherence-based agent architecture along with the reasoning algorithm.

## 5.2 Adaptation of MCS

In this section, we first give an overview of the adaptation of MCS specification of an agent followed by the details on each of the contexts and the reasoning mechanism across contexts. We adapt the MCS specification of g-BDI architecture. As briefly mentioned in Section 2.2, the MCS specification of an agent contains three basic components: units or contexts, logics, and bridge rules that channel the propagation of consequences between theories. An agent in this architecture may be defined as a tuple $\langle \{C_i\}_{i=1\ldots n}, B\rangle$ consisting of:

- a family $\{C_i\}_{i=1\ldots n}$ of contexts, $n > 0$, where each context $C_i = \langle L_i, A_i, \vdash_i, \mathcal{T}_i\rangle$ consists of a language $L_i$, a set of axioms $A_i$, and an MDR $\vdash_i$ defining the logical system, together with a theory presentation $\mathcal{T}_i \subseteq L_i$ of the context.
- a set $B$ of bridge rules, i.e., inference rules of the form

$$\frac{i_1 : A_1 \quad i_2 : A_2 \quad \cdots \quad i_q : A_q}{j : A}$$

  where $i_k$ (with $k \in \{1, \ldots, q\}$ and $q > 0$) and $j$ are indeces of contexts (i.e., $1 \leq i_k, j \leq n$), and $A_k$ and $A$ are formula schemata specifying premises from contexts $C_{i_k}$ and a conclusion from context $C_j$, respectively. (Later we extend the notion of bridge rules to cope with graded formulas as introduced below.)

In our adaptation of the multi-context architecture, the theories $\mathcal{T}_i$ of the contexts will yield coherence graphs. We have already defined the coherence function $\zeta$ derived from an MDR $\vdash_i$ within one context (see Definition 3.2.5).

In the following we define how this coherence function is to be extended to capture coherence that arises between formulas due to bridge rules carrying consequences from one graph to another. For this we will define two additional kinds of functions, graph-extension and graph-join functions. First we begin by giving a brief overview of the contexts in our multi-context system before defining these functions.

### 5.2.1 Cognitive Contexts

We discuss the belief, desire, and intention contexts of an intentional agent with graded cognition. We take the belief, desire, and intention contexts as defined in Casali et al. [Casali et al., 2005]. We here give a sketch of a belief context, while the details are in Casali et al. [Casali et al., 2006]. The desire and intention contexts can be defined in a similar fashion, with the belief logic replaced either by a desire or intention logic accordingly. Further, the belief theory $\mathcal{T}_B$ gives rise to a coherence graph whose nodes are graded formulas of the belief language. We call this graph a *belief graph* which is realised by extending the deductive coherence function.

#### Belief Logic

Following Casali et al., a belief logic $\mathcal{K}_B = \langle L_B, A_B, \vdash_B \rangle$ consists of a belief language $L_B$, a set of axioms $A_B$ and an MDR $\vdash_B$. We define the belief language $L_B$ by extending a classical propositional language $L$ defined upon a countable set of propositional variables and connectives $\neg$ and $\rightarrow$, with a fuzzy unary modal operator $B$. The modal language $L_B$ is built from elementary modal formulas $B\varphi$ (where $\varphi$ is propositional) and truth constants $\bar{r}$ (for each rational $r \in \mathbb{Q} \cap [0,1]$) using the connectives of Łukasiewicz many-valued logic. If $\varphi$ is a proposition in $L$, the intended meaning of $B\varphi$ is that "$\varphi$ is believable". We use a modal many-valued logic based on Łukasiewicz logic to formalise $\mathcal{K}_B$ as follows: [1]

1. Given a propositional language $L$, the belief language $L_B$ of $\mathcal{K}_B$ is given by:

   - If $\varphi \in L$ then $B\varphi \in L_B$
   - If $r \in \mathbb{Q} \cap [0,1]$ then $\bar{r} \in L_B$
   - If $\Phi, \Psi \in L_B$ then $\Phi \rightarrow_L \Psi \in L_B$ and $\Phi \,\&\, \Psi \in L_B$ (where $\&$ and $\rightarrow_L$ correspond to conjunction and implication of Łukasiewicz logic)

   Other Łukasiewicz logic connectives for the modal formulas can be defined from $\&$, $\rightarrow_L$ and $\overline{0}$: $\neg_L\Phi$ is defined as $\Phi \rightarrow_L \overline{0}$, $\Phi \wedge \Psi$ as $\Phi \,\&\, (\Phi \rightarrow_L \Psi)$, $\Phi \vee \Psi$ as $\neg_L(\neg_L\Phi \wedge \neg_L\Psi)$, and $\Phi \equiv \Psi$ as $(\Phi \rightarrow_L \Psi) \,\&\, (\Psi \rightarrow_L \Phi)$.

2. The axioms $A_B$ of $\mathcal{K}_B$ are:

[1] We could use other logics as well by replacing the axioms.

- all axioms of propositional logic;
- the axioms of Łukasiewicz logic for modal formulas (for instance, axioms of Hájek's Basic Logic (BL) [Hájek, 1998] plus the axiom $\neg_{BL}\neg_{BL}\Phi \rightarrow_{BL} \Phi$);
- the probabilistic axioms, i.e., given $\varphi, \psi \in L$,

$$\begin{aligned} B(\varphi \rightarrow \psi) &\rightarrow_L (B\varphi \rightarrow_L B\psi) \\ B\varphi &\equiv \neg_L B(\varphi \wedge \neg\psi) \rightarrow_L B(\varphi \wedge \psi) \\ \neg_L B\varphi &\equiv B\neg\varphi \end{aligned}$$

3. Finally, the MDR $\vdash_B$ of $\mathcal{K}_B$ is defined by the inference rules of
   - modus ponens;
   - necessitation for $B$ (i.e., from $\varphi$ derive $B\varphi$).

Since in Łukasiewicz logic a formula $\Phi \rightarrow_L \Psi$ is 1-true if, and only if, the truth value of $\Psi$ is greater or equal to that of $\Phi$, modal formulas of the type $\bar{r} \rightarrow_L B\varphi$ express that the probability of $B\varphi$ is at least $r$. We shall use the notation $(B\varphi, r)$ for these kind of formulas, and call them *graded beliefs*. Let $L_B^* \subseteq L_B$ denote the set of all graded beliefs of $L_B$. Furthermore, we shall only consider theory presentations $\mathcal{T}_B \subseteq L_B^*$ expressed using graded beliefs.

### Belief Graph

Our aim is, given a theory presentation $\mathcal{T}_B \subseteq L_B^*$ expressed using graded beliefs in a belief language, to define the corresponding coherence graph (see Definition 5.2.4 further below). For this, however, we first need to extend the definitions given in Chapter 3 for graded beliefs underling a belief language $L_B$. The idea is to determine coherence values between graded beliefs not only by virtue of the deduction relation; we will also take into account the grades as specified in the theory presentation $\mathcal{T}_B \subseteq L_B^*$.

**Definition 5.2.1** *Let $L_B$ be the belief language as defined above, and let $\mathcal{T}_B \subseteq L_B^*$ be a finite theory presentation using only graded formulas. We call $\mathcal{T}_B^\bullet$ the* closure of $\mathcal{T}_B$ under subformulas *when $(B\varphi', r') \in \mathcal{T}_B^\bullet$ if and only if there is an $(B\varphi, r) \in \mathcal{T}_B$ such that $\varphi'$ is a subformula of $\varphi$ and $r' = \sup\{q \mid \mathcal{T}_B \models (B\varphi', q)\}$.*

**Definition 5.2.2** *Let $L_B$ be the belief language and $\vdash_B$ the MDR as defined above. Let $\mathcal{T}_B \subseteq L_B^*$ be a finite theory presentation using only graded formulas.*

*A* support function $\eta_{\mathcal{T}_B} : \mathcal{T}_B^\bullet \times \mathcal{T}_B^\bullet \to [-1,1]$ *with respect to* $\mathcal{T}_B$ *is given by:*

$$\eta_{\mathcal{T}_B}(\Phi,\Psi) = \begin{cases} \max \begin{cases} \frac{r}{|\Gamma|+1} & \text{where } \Gamma \text{ is the smallest subset of } \mathcal{T}_B^\bullet \text{ such that} \\ & \Gamma, \Phi \vdash_B \Psi \text{ and } \Gamma \nvdash_B \Psi \text{ and } \Phi \nvdash_B \text{ and } \Psi = (\alpha, r) \\ \frac{r}{|\Gamma|+2} & \text{where } \Gamma \text{ is the smallest subset of } \mathcal{T}_B^\bullet \text{ such that} \\ & \exists (\alpha, r) \in \mathcal{T}_B^\bullet \text{ with } \alpha \neq \bar{0} \text{ such that} \\ & \Gamma, \Phi, \Psi \vdash_B (\alpha, r) \text{ and } \Gamma, \Phi \nvdash_B (\alpha, r) \text{ and} \\ & \Gamma, \Psi \nvdash_B (\alpha, r) \text{ and} \\ \frac{-r}{|\Gamma|+1} & \text{where } \Gamma \text{ is the smallest subset of } \mathcal{T}_B^\bullet \text{ such that} \\ & \Gamma, \Phi, \Psi \vdash_B (\bar{0}, r) \text{ and } \Gamma, \Phi \nvdash_B (\bar{0}, r) \\ & \text{and } \Gamma, \Psi \nvdash_B (\bar{0}, r) \end{cases} \\ \text{undefined, otherwise} \end{cases}$$

**Definition 5.2.3** *Let* $L_B$ *be the belief language as defined above. Let* $\mathcal{T}_B \subseteq L_B^*$ *be a finite theory presentation using only graded formulas. Let* $\eta_{\mathcal{T}_B} : \mathcal{T}_B^\bullet \times \mathcal{T}_B^\bullet \to [-1,1] \setminus \{0\}$ *be a support function with respect to* $\mathcal{T}_B$*. A deductive coherence function* $\zeta_{\mathcal{T}_B} : (\mathcal{T}_B^\bullet)^{(2)} \to [-1,1] \setminus \{0\}$ *with respect to* $\mathcal{T}_B$ *is given by:*

$$\zeta_{\mathcal{T}_B}(\{\Phi,\Psi\}) = \begin{cases} \max\{\eta_{\mathcal{T}_B}(\Phi,\Psi), \eta_{\mathcal{T}_B}(\Phi,\Psi)\} & \text{if } \eta_{\mathcal{T}_B}(\Phi,\Psi) \neq 0 \text{ and } \eta_{\mathcal{T}_B}(\Psi,\Phi) \neq 0 \\ \eta_{\mathcal{T}_B}(\Phi,\Psi) & \text{if } \eta_{\mathcal{T}_B}(\Phi,\Psi) \neq 0 \\ & \text{and } \eta_{\mathcal{T}_B}(\Psi,\Phi) = 0 \text{ or is undefined} \\ \text{undefined} & \text{if } \eta_{\mathcal{T}_B}(\Phi,\Psi) = 0 \text{ or is undefined} \\ & \text{and } \eta_{\mathcal{T}_B}(\Psi,\Phi) = 0 \text{ or is undefined} \end{cases}$$

**Definition 5.2.4** *Let* $\mathcal{K}_B = \langle L_B, A_B, \vdash_B \rangle$ *be a belief logic, where* $L_B$ *is a belief language,* $A_B$ *are a set of axioms and* $\vdash_B$ *is an MDR. Let* $\mathcal{T}_B \subseteq L_B^*$ *be a finite theory presentation expressed using graded beliefs only. A* belief graph *of* $\mathcal{T}_B$ *is the coherence graph* $g = \langle V, E, \zeta \rangle$ *where*

- $V = \mathcal{T}_B^\bullet$
- $\zeta = \zeta_{\mathcal{T}_B}$
- $E = \{\{\Phi,\Psi\} \in V^{(2)} \mid \zeta_{\mathcal{T}_B}(\{\Phi,\Psi\}) \text{ is defined}\}$

A belief graph represents the graded beliefs of an agent (and all the subformulas thereof) and the coherences and incoherences among them. A desire graph and intention graph of theory presentations $\mathcal{T}_D$ and $\mathcal{T}_I$ in logics $\mathcal{K}_D$, and $\mathcal{K}_I$, respectively, would be similarly defined.

### 5.2.2 Reasoning Across Contexts

Reasoning in a BDI agent needs to consider the influence of cognitions among each other. For instance, it is desirable to choose or predict an action that is most

coherent with the set of cognitions. It is also desirable to know the influence of a new information on the overall coherence of cognitions. Typically, in a multi-context system, such reasoning is achieved by the use of bridge rules. For coherence-driven agents we adapt the idea of bridge rules to be able to establish links and coherence values across several coherence graphs.

Bridge rules are in a certain sense inference rules carrying inferences between theories of different logics. Since theories determine the nodes of coherence graphs, we can use these bridge rules to find coherence values (and thus edges) between nodes of different graphs. However, we generalise this process to include any inference rules which take premises and conclusion from theories of different contexts. For this, we define two kinds of functions on tuples $\bar{g} = \langle g_1, \ldots, g_n \rangle$ in $\mathcal{G}^n$ representing the coherence graphs determined by a collection of theory presentations of various contexts.

First, we shall define functions that extend individual coherence graphs $g_i$ with new nodes whenever the corresponding formulas can be derived using inferences across contexts. In these cases there will exist a positive coherence relation between the premises and the conclusion of context-bridging inference rules. Consequently, we also define functions that make the union of the all coherence graphs $g_i$ and further add those edges between nodes coming from premises and conclusions of context-bridging inference rules. Below we define both the graph-extension and edge-join functions and, finally, we discuss the definition and application of these functions for bridge rules.

A graph-extension function (denoted with $\varepsilon$) takes into account the influence of graphs on each other. (For example, when an agent wants it to be the case that, whenever it has an intention $(I\varphi, r)$ in the intention graph, then a corresponding belief $(B\varphi, r)$ is inferred into the belief graph.)

**Definition 5.2.5** *We say that a function $\varepsilon : \mathcal{G}^n \to \mathcal{G}^n$, $n > 0$, is a* graph-extension function *if, given a tuple of graphs $\bar{g} = \langle g_1, \ldots, g_n \rangle$ in $\mathcal{G}^n$, $\varepsilon(\bar{g}) = \bar{g}'$ is such that*

- $V_i' \supseteq V_i$
- $E_i' = E_i$
- $\zeta_i' = \zeta_i$.

*Let $\mathcal{E}$ denote the set of all graph-extension functions (for a fixed $n$).*

A desirable property for a function $\varepsilon \in \mathcal{E}$ would be to have fixed points. This is because a fixed point gives us a terminating condition for the repeated application of an extension function. We call a tuple of graphs $\bar{g}$ a *fixed point* of a subset $S \subseteq \mathcal{E}$, if the value of the application of any extension function in $S$ on $\bar{g}$ is $\bar{g}$.

**Definition 5.2.6** *We say that a sequence is an* extension sequence *if, given a tuple of graphs $\bar{g} \in \mathcal{G}^n$, $n > 0$, and a set of graph-extension functions $S \subseteq \mathcal{E}$,*

$$g^0 = \{\bar{g}\}, \ldots, g^i = \{\varepsilon(\bar{h}) \mid \bar{h} \in g^{i-1} \wedge \varepsilon \in S\}, \ldots$$

*and say that the elements of* $g^j$, $j > 0$, *are* fixed points *of* $S$ *applied over* $\bar{g}$ *(denoted as* $S^*(\bar{g})$*) if* $g^j = g^{j-1}$. *Further, we say that the fixed point is unique if* $|S^*(\bar{g})| = 1$.

A graph-join function (denoted with $\iota$) takes $n$ graphs and joins them together, further adding new edges (and coherence values on the edges) between certain nodes. This does not change the theories, as this function only makes new associations between formulas of different theories. (For example, when an agent wants it to be the case that, whenever an intention $(I\varphi, r)$ in the intention graph has lead to a corresponding belief $(B\varphi, r)$ in the belief graph, these two formulas are to be related by positive coherence.)

**Definition 5.2.7** *We say that a function* $\iota : \mathcal{G}^n \to \mathcal{G}$, $n > 0$, *is a* graph-join function *if, given a tuple of graphs* $\bar{g} = \langle g_1, \ldots, g_n \rangle$ *in* $\mathcal{G}^n$, *with* $g_i = \langle V_i, E_i, \zeta_i \rangle$, $\iota(\bar{g}) = \langle V, E, \zeta \rangle$ *is such that:*

- $V = \bigcup_{1 \leq i \leq n} \{i : \Phi \mid \Phi \in V_i\}$
- $E \supseteq \bigcup_{1 \leq i \leq n} \Big\{ \{i : \Phi, i : \Psi\} \mid \{\Phi, \Psi\} \in E_i \Big\}$
- $\zeta : E \to [-1, 1] \setminus \{0\}$ such that $\zeta(\{i : \Phi, i : \Psi\}) = \zeta_i(\{\Phi, \Psi\})$

*Let* $\mathcal{J}$ *denote the set of all graph-join functions (for a fixed* $n$*).*

Now we define the composition of graphs in a tuple $\bar{g}$ by combining the two kinds of functions. That is, we apply a graph-join function $\iota \in \mathcal{J}$ on the fixed point of a set of graph-extension functions $S \subseteq \mathcal{E}$ applied over $\bar{g}$. Note that here we assume $S$ has a unique fixed point applied over any tuple of graphs $\bar{g}$. This is, however, a fair assumption given that we can construct the functions in $S$ according to the requirements. Further, it should be noted that we keep the theories separate and only compose the corresponding coherence graphs.

**Definition 5.2.8** *We say that a function* $\varsigma : \mathcal{G}^n \to \mathcal{G}$, $n > 0$, *is a* graph-composition function *if, given a tuple of graphs* $\bar{g}$ *in* $\mathcal{G}^n$, *a set of graph-extension functions* $S \subseteq \mathcal{E}$ *with a unique fixed point and a graph-join function* $\iota \in \mathcal{J}$, $\varsigma(\bar{g}) = \iota(S^*(\bar{g}))$.

### Bridge Rules — Composition Functions

Now we describe how such graph-composition functions can be derived from bridge rules. Bridge rules have been used in multi-context systems to make inferences across contexts. Here we use them to derive coherence associations across graphs that correspond to theory presentations of graded formulas.

**Definition 5.2.9** *Given a family $\{C_i\}_{i=1\ldots n}$ of contexts, $n > 0$, a bridge rule $b$ is a rule of the form*

$$\frac{i_1 : (A_1, R_1) \quad i_2 : (A_2, R_2) \quad \cdots \quad i_q : (A_q, R_q)}{j : (A, f(R_1, R_2, \ldots, R_q))}$$

*where:*

- *$i_k$ (with $k \in \{1, \ldots, q\}$ and $q > 0$) and $j$ are all pairwise distinct[2] indeces of contexts (i.e., $1 \leq i_k, j \leq n$)*
- *$A_k$ and $A$ are are formula schemata specifying premises from contexts $C_{i_k}$ and a conclusion from context $C_j$, respectively*
- *$R_k$ are either variables or numerical constants in $\mathbb{Q} \cap [0, 1]$*
- *$f(R_1, R_2, \ldots, R_q)$ is an expression, where $f : (\mathbb{Q} \cap [0, 1])^q \to \mathbb{Q} \cap [0, 1]$*

*Let $\mathcal{B}$ denote the set of all such bridge rules.*

Given a bridge rule $b \in \mathcal{B}$, we derive a graph-extension function that, given tuple $\bar{g} = \langle g_1, \ldots, g_n \rangle$ of graphs, extends graph $g_j$ with a new node corresponding to an instance of the conclusion schema $A$.

**Definition 5.2.10** *Let $\{C_i\}_{i=1\ldots n}$ be a family of contexts, $n > 0$, and let*

$$b = \frac{i_1 : (A_1, R_1) \quad i_2 : (A_2, R_2) \quad \cdots \quad i_q : (A_q, R_q)}{j : (A, f(R_1, R_2, \ldots, R_q))}$$

*be a a bridge rule as in Definition 5.2.9. The graph-extension function $\varepsilon_b$ is defined as follows: Given a tuple of graphs $\bar{g} = \langle g_1, \ldots, g_j, \ldots, g_n \rangle$ (with $g_j = \langle V_j, E_j, \zeta_j \rangle$), then $\varepsilon_b(\bar{g}) = \langle g_1, \ldots, g'_j, \ldots, g_n \rangle$ where $g'_j = \langle V'_j, E'_j, \zeta'_j \rangle$ such that*

- *$V'_j = V_j \cup \{(\pi(A), f(\pi(R_1), \pi(R_2), \ldots, \pi(R_q)))\}$ if there exists a most general substitution $\pi$ such that, for all $k \in \{1, \ldots, q\}$, $(\pi(A_k), \pi(R_k)) \in V_k$; otherwise $V'_j = V_j$*
- *$E'_j = E_j$*
- *$\zeta'_j(\{v, w\}) = \zeta_j(\{v, w\})$ for all $v, w \in V_j$*

Given a finite set of bridge rules $B \subseteq \mathcal{B}$, we derive a graph-join function that, given tuple $\bar{g} = \langle g_1, \ldots, g_n \rangle$ of graphs, joins all graphs $g_i$ together, adding new edges and their coherence values between nodes corresponding to instances of premise schemata $A_k$ and the conclusion schema $A$, and also between instances of premise schemata themselves, in accordance to Thagard's principles discussed in Section 3.2.1.

[2]This condition can be dispensed with; we only require it for subsequent ease of presentation of Definition 5.2.11 below.

**Definition 5.2.11** *Let $\{C_i\}_{i=1\ldots n}$ be a family of contexts, $n > 0$, and let $B \subseteq \mathcal{B}$ be a finite subset of bridge rules. The graph-join function $\iota_B$ is defined as follows: Given a tuple of graphs $\bar{g} = \langle g_1, \ldots, g_n\rangle$ (with $g_i = \langle V_i, E_i, \zeta_i\rangle$), then $\iota_B(\bar{g}) = \langle V, E, \zeta\rangle$ such that*

- $V = \bigcup_{1\leq i\leq n} \{i : \Phi \mid \Phi \in V_i\}$
- $E = \bigcup_{1\leq i\leq n} \Big\{\{i : \Phi, i : \Psi\} \mid \{\Phi, \Psi\} \in E_i\Big\} \cup$
$$\bigcup_{b\in B} \left\{ \{i : \Phi, j : \Psi\} \;\middle|\; \begin{array}{l} i : \Phi \text{ is a premise of } \pi(b) \text{ and } j : \Psi \text{ is the conclusion of } \pi(b), \\ \text{where } \pi \text{ is a most general substitution, such that,} \\ \text{for all premises } k : (A, R) \text{ of } b,\ \pi((A, R)) \in V_k \end{array} \right\} \cup$$
$$\bigcup_{b\in B} \left\{ \{i : \Phi, j : \Psi\} \;\middle|\; \begin{array}{l} i : \Phi \text{ and } j : \Psi \text{ are premises of } \pi(b),\ i \neq j, \\ \text{where } \pi \text{ is a most general substitution, such that,} \\ \text{for all premises } k : (A, R) \text{ of } b,\ \pi((A, R)) \in V_k \end{array} \right\}$$
- $\zeta(\{i : \Phi, i : \Psi\}) = \zeta_i(\{\Phi, \Psi\})$, *and* $\zeta(\{i : \Phi, j : \Psi\})$ *for $i \neq j$ is defined as in Definition 5.2.3 with respect to the following support function:*

$$\eta(i : \Phi, j : \Psi) = \begin{cases} \max \left\{ \begin{array}{ll} \frac{r}{|\Gamma|+1} & \text{where } \Gamma \text{ is the smallest subset of } V, \text{ such that} \\ & \exists b \in B \text{ such that } \Gamma \cup \{i : \Phi\} \text{ is the set of premises} \\ & \text{and } j : \Psi \text{ with } \Psi = (\alpha, r) \text{ is the conclusion of } \pi(b), \\ & \text{where } \pi \text{ is a most general substitution, such that,} \\ & \text{for all premises } k : (A, R) \text{ of } b,\ \pi((A, R)) \in V_k \\ \frac{r}{|\Gamma|+2} & \text{where } \Gamma \text{ is the smallest subset of } V, \text{ such that} \\ & \exists b \in B \text{ such that } \Gamma \cup \{i : \Phi, j : \Psi\} \text{ is the set of} \\ & \text{premises and } h : (\alpha, r) \text{ is the conclusion of } \pi(b), \\ & \text{where } \pi \text{ is a most general substitution, such that,} \\ & \text{for all premises } k : (A, R) \text{ of } b,\ \pi((A, R)) \in V_k \end{array} \right\} \\ \text{undefined, otherwise} \end{cases}$$

**Application of Composition Functions — An Example**

Consider, for example, the tuple of graphs $\langle g_{\mathrm{B}}, g_{\mathrm{D}}, g_{\mathrm{I}}\rangle$ corresponding to a multi-context system with belief context $C_{\mathrm{B}}$, desire context $C_{\mathrm{D}}$ and intention context $C_{\mathrm{I}}$, and with a single bridge rule

$$b = \frac{\mathrm{B} : (B\varphi, r) \quad \mathrm{D} : (D\varphi, s)}{\mathrm{I} : (I\varphi, \min(r, s))}$$

Let's further assume that $(Bp, 0.95)$ is a node in $g_{\mathrm{B}}$ and $(Dp, 0.7)$ is a node in $g_{\mathrm{D}}$ where $p$ is a proposition. Then,

- $\varepsilon_b(\langle g_{\mathrm{B}}, g_{\mathrm{D}}, g_{\mathrm{I}}\rangle) = \langle g_{\mathrm{B}}, g_{\mathrm{D}}, g'_{\mathrm{I}}\rangle$, where $g'_{\mathrm{I}}$ is $g_{\mathrm{I}}$ with the node $(Ip, 0.7)$ added to its set of nodes.

- $\iota_{\{b\}}(\langle g_{\mathrm{B}}, g_{\mathrm{D}}, g'_{\mathrm{I}}\rangle)$ is the disjoint union of graphs $g_{\mathrm{B}}$, $g_{\mathrm{D}}$, and $g'_{\mathrm{I}}$ with the additional edges $\{\mathrm{B} : (Bp, 0.95), \mathrm{I} : (Ip, 0.7)\}$, $\{\mathrm{D} : (Dp, 0.7), \mathrm{I} : (Ip, 0.7)\}$, and $\{\mathrm{B} : (Bp, 0.95), \mathrm{D} : (Dp, 0.7)\}$, with coherence value 0.35 for all of them.

## 5.3 Coherence-Driven Agents

Equipped with contexts and a mechanism to reason across these contexts, we can now formally define a coherence-driven agent. Recall that, the MCS specification of an agent is a group of interconnected contexts $\langle\{C_i\}, B\rangle$. Each context is a tuple $C_i = \langle L_i, A_i, \vdash_i, \mathcal{T}_i\rangle$ where $L_i$, $A_i$ and $\vdash_i$ are the language, axioms, and inference rules of a logical system, and $\mathcal{T}_i$ is a finite theory presentation. In our extension of MCS, a coherence-driven agent will further have a function cohgraph that maps a context to its corresponding coherence graphs. And a function compfun that maps a set of bridge rules to a graph-composition function. This extension is required because contexts are expressed as coherence graphs and agents will need both coherence and graph-composition functions to reason within and between graphs. For the normative BDI agents considered here, the contexts are $C_{\mathrm{B}}$, $C_{\mathrm{D}}$ and $C_{\mathrm{I}}$, which determine a belief graph $g_{\mathrm{B}}$, a desire graph $g_{\mathrm{D}}$, and an intention graph $g_{\mathrm{I}}$ respectively. Hence we have the following definition:

**Definition 5.3.1** *A coherence-driven agent $a$ is a tuple $\langle\{C_i\}_{i=\mathrm{B,D,I,N}}, B, \mathsf{cohgraph}, \mathsf{compfun}\rangle$ where $\{C_i\}_{i=\mathrm{B,D,I}}$ is a family of contexts, $B \subseteq \mathcal{B}$ is a set of bridge rules, $\mathsf{cohgraph} : \{C_i\}_{i=\mathrm{B,D,I}} \to \mathcal{G}$ maps contexts to coherence graphs, and $\mathsf{compfun} : 2^B \to \mathcal{G}^{(\mathcal{G}^3)}$ maps sets of bridge rules to composition functions that take a triple of graphs to a graph.*

In the following we describe how coherence-driven agents interact with an external environment. As in Figure 5.1, a coherence-driven agent starts with a set $\{C_i\}_{i=\mathrm{B,D,I}}$ of contexts corresponding to the beliefs, desires, and intentions, which it expresses as coherence graphs. We assume that the languages of each context are all extensions of the same propositional language $L$. It is also desirable for the theory presentations of contexts to be consistent. Our proposal, however, is tolerant to inconsistencies and in a certain sense exposes them and eliminates them, if possible. Further, the agent is assumed to have a set $B$ of bridge rules to reason across contexts and it computes the composite graph using these.

An agent at any time can either perceive the external environment or make a prediction about a future action. In the event of a new piece of information $(K\varphi, r)$ (where $K$ is one of the modal operator of its context languages) the agent reasoning proceeds according to the following algorithm:

1. it adds the new graded formula to the theory $\mathcal{T}$ of the corresponding context $C_{\mathrm{B}}$, $C_{\mathrm{D}}$ or $C_{\mathrm{I}}$ (depending if $K$ is either $B$, $D$, $I$ respectively);

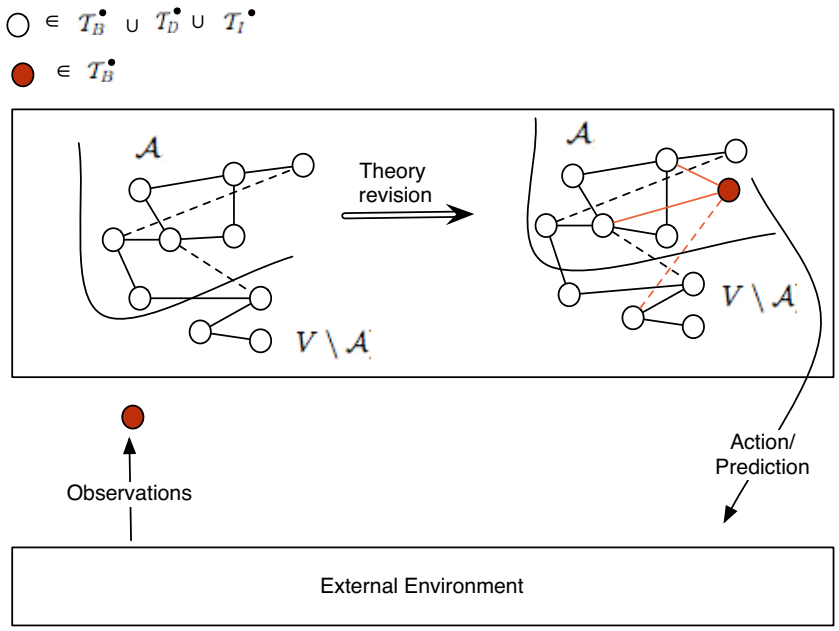


Figure 5.1: Architecture of a coherence-driven agent

2. it computes the deductive closure of $\mathcal{T}$ (but, to keep the closure finite, we compute it limited to $\mathcal{T}^\bullet$, i.e., without introducing formulas with new subformulas that are not in $\mathcal{T}$);

3. it expresses the contexts with newly closed theories as coherence graphs and computing the tuple $\bar{g} = \langle \mathsf{cohgraph}(C_\mathrm{B}), \mathsf{cohgraph}(C_\mathrm{D}), \mathsf{cohgraph}(C_\mathrm{I}) \rangle$ associated to them;

4. it computes the composite graph $g = \mathsf{compfun}(\bar{g}) = \iota_B(S^*(\bar{g}))$, where $S = \{\varepsilon_b \mid b \in B\}$;

5. it computes all coherence-maximising partitions[3] $(\mathcal{A}, V \setminus \mathcal{A})$, where $V$ is the set of nodes of $g$, by computing $\arg\max_{\mathcal{A} \subseteq V} \sigma(g, \mathcal{A})$, and eventually chooses one of them;[4]

[3] Finding a maximising partition of an edge-weighted graph is known to be an NP-complete problem. There exist algorithms computing an approximation to the solution to this problem, such as max-cut or neural-network based algorithms.

[4] According to the guidelines discussed in Chapter 3, we can decide on a favourable accepted set. However it should also be remembered that, coherence maximisation is more about understanding which pieces of information can be accepted together rather than providing an ultimate answer to which piece of information should be accepted.

6. it updates the theory presentations of the contexts according to the newly accepted set.

As discussed in Chapter 3, if the new piece of information $(K\varphi, r)$ reinforces the original theory, it is added to the accepted set and the theory becomes more coherent. If it contradicts elements of the original theory, then either the new piece of information is rejected, or some part of the already accepted theory is rejected, whichever makes the theory more coherent. To make predictions, however, the agent uses only the accepted theory. This is realistic, as it is the accepted set that the agent wants to base its decisions on.

Another important observation is regarding the values of function $\sigma$. In theory, the coherence of the graph $\kappa(g)$ is set as the maximum of the strength values $\sigma(g, \mathcal{A})$; in reality, this could be very much dependent on the agent. If the inclusion of a node only slightly reduces the coherence of the graph, a mildly distressed agent may choose to ignore the incoherence, or may be satisfied with modifying the degree on the node. Whereas a heavily distressed agent may not only choose to reject the corresponding cognition, but may actively seek out information or action to change the cognitions.

## 5.4 Discussion

This chapter defines a coherence-based agent architecture based on the framework defined in the previous chapters. The architecture is based on the BDI architecture for agents which models the mental attitudes of the agent. Further, the architecture takes into account that agents model of the world is incomplete and imprecise and hence we use graded version of the above architecture (g-BDI architecture). It further extends the architecture to make coherence as the driving mechanism by structuring theories in each of the modalities as coherence graphs and defining a way to combine them and reason across them.

Further, an algorithm for coherence-driven reasoning in such agents is sketched which emphasis the main philosophy behind a coherence-driven agent, that is, the reasoning in such an agent is driven by the maximisation of coherence. In the following chapter, we will see that this can in fact make the agents very flexible and autonomous decision maker under dynamic situations.

# Chapter 6

# Coherence-driven Agents—Experimental Evaluation

In this chapter, we experimentally evaluate the performance of a coherence-driven agent and test the feasibility of our approach. For the purpose, we have designed a game scenario where a player has to choose autonomously between possible actions in a dynamic environment. We then compare the performances of different types of players (human, coherence-driven agent, and near optimal agent— agents that are tuned to perform near optimally for this specific experiment). We choose this particular experiment primarily because we only aim to demonstrate the feasibility of our approach and hence needed a very simple scenario. Nevertheless, the experiment has the potential to grow in complexity since action selection in autonomous agents situated in dynamic environments is a complex problem requiring much flexibility and adaptability in agent architecture. For the current purposes however, we limit the experiment to a simple action selection problem in a controlled environment.

In Section 6.1, we introduce the experimental set-up, the hypothesis and variables used to control the experiment. Section 6.2 discusses the three types of agents used in the experiment. Simulation results are discussed in Section 6.3. We conclude with discussions in Section 6.4.

## 6.1 Experimental Evaluation

In this section, we discuss the experimental set-up, the hypotheses, variables that control the experiment, and the evaluation criteria.

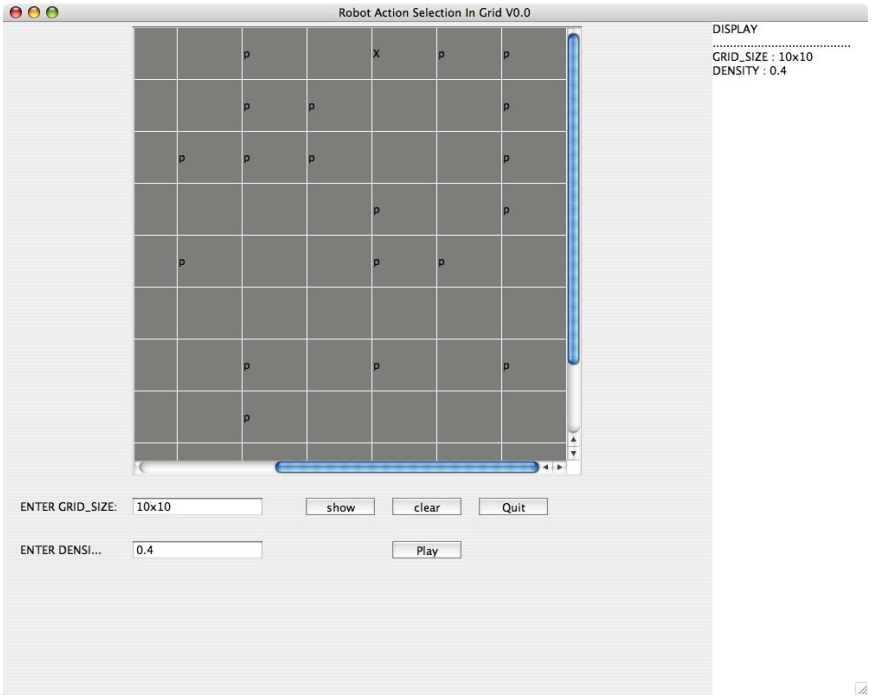


Figure 6.1: Grid Environment

### 6.1.1 Experimental Set-up

The experiment is setup in a two-dimensional board as in Figure 6.1 where a player can move around the cells. There are two kinds of cells *ordinary* and *plug enabled*, the latter having a provision to "plug". The simulation is run for a fixed number of rounds, and at each round a player can choose between two possible actions: "plug" to restore energy or "move" to earn points. A player can only move to any of its immediate neighbouring cells and the cost of each move is a reduction in energy level by a fixed quantity. Each plug action restores the energy levels to the maximum. The goal of a player placed on the board environment is to explore the board. Hence, there is a reward for every move action, which is a fixed number of points $n(\geq 1)$, the reward for a plug action is 0 points. The density of plugs on the board can be set to a value between 0 and the number of cells on the board and may be varied during a run. The game ends when either all rounds are completed or if the remaining energy becomes 0. The performance of a player is evaluated by the number of points earned during the total run length of the simulation. The more moves a player makes, the better its performance. However, depending on the density of the plugs distributed in the environment and the remaining energy level of the player, it has to make a clever choice between moving and charging in order to stay alive.

### 6.1.2 Variables

The variables used in the experiment and the range of values it can assume are given in Table 6.1.

| **Name** | **Range** | **Selected Values** |
|---|---|---|
| Player type | any | *human, coherence, near-opt* |
| Run length (*Rounds)* | $\mathbb{N} > 1$ | 100 |
| Plug Distribution | | random |
| Density change frequency | [0, 100]% | 0, 20, 50, 80 |
| Density Values | [0, 100]% | randomly chosen within [20,80]% |
| Plug positions | Any set of values in [0, *no. of cells]* | randomly drawn sets of values in the range [0, *no. of cells]* set size equal to current density |
| Cost of a *move* action | $0 \leq cost \leq 1$ | 0.1 |
| Reward for a *move* action | $\leq 0$ | 1 |
| Cost of a *plug* action | $0 \leq cost \leq 1$ | 0 |
| Reward for a *plug* action | $\leq 0$ | 0 |

Table 6.1: Simulation variables

*Player Type:* We have considered three different types of players for the experiment. An agent player which has a coherence-driven decision making module (*coherence*), human players chosen from a randomly generated population (*human*), and another agent player, which has a near-optimal algorithm tuned for playing the particular game (*near-opt*). The information on the round the player currently is in, the density distribution of plugs, the remaining charge and the points earned so far are available for the players at each round in the simulation. To detect these values, a player is equipped with an *energy_sensor* ($e_s$), a *plug_sensor* ($p_s$) and a *density_sensor* ($d_s$). Sensor $e_s$ senses the energy level of the player where as $p_s$ tells the player whether it can plug at the cell it is in. The density sensor $d_s$ finds the current density of plug distribution, however, the positions of actual plugs are not known to any of the players. The information on whether a player can plug at a cell is known only when it is on that cell.

*Run Length:* Run length represents the length of a single run in a game and is measured in number of rounds played. The run length for the current simulations is fixed at 100 rounds. We do not vary run length for the simulation because the effect of run length is dependent on the change in plug density. That is, if the density is high, no matter the number of runs, the players complete the runs. Whereas in a low density, again, irrespective of the run length players performance decline. Hence, by varying density of the plug distribution, we emulate the run length. *Density Change Frequency:* The values assigned are in

percentage $0, 20, 50$ and $80$. Here $0\%$ means that the density once assigned will not be changed for the entire run of the experiment, where as a value of $80\%$ means that the density value and hence the plug distribution will be reassigned $80\%$ of the time during a single simulation run. That is, at each round there is an $80\%$ probability that the density and hence the plug distribution is reassigned.

*Density Values & Plug Distribution:* During the run length, when a density value needs to be changed, they are assigned a random value between $[20, 80]\%$. Initial experimentation has prompted us to omit very low density values below $20\%$ and very high values above $80\%$. This is because, irrespective of other variable values, all players perform equally badly at very low density values and equally well at very high density values. Since we assign density values randomly, allowing values between $[0, 20)\%$ and $(80, 100]\%$ will have unwanted influence on the experimental result and subsequent conclusions. The plug distribution for any particular density also is assigned randomly, drawn from sets of values within the range of the board.

*Cost & Reward for Actions:* The cost of a move action is fixed at 0.1 reduction in energy levels and reward of a move action is fixed at 1 point per move. These could be varied to study the effect of accelerating cost and reward respectively. However, from initial experiments, we do not anticipate that these variables are significant in determining the result. Similarly, cost and reward of a *plug* action is fixed at 0.

### 6.1.3 Hypothesis

The research objective stated in Chapter 1 was to design artificial agents that are more autonomous by making them more flexible and adaptive.

> Can we define a suitable framework to model autonomous reasoning of agents which can incorporate uncertainty and dynamism in the agent's world model while not loosing the type of formal qualities such as testability and reliability?

It is hard to demonstrate whether autonomy actually works unless the problem domain is complex enough for non-autonomous agents to stumble. However, the scenario we set up is simple enough to design a near optimal algorithm. Hence, we will not be able to validate experimentally the claims that coherence-driven agents are more flexible and adaptive and can perform better than rigid algorithms tuned to solve a specific problem. Hence, we conduct these experiments with the aim that the performance of a coherence-driven agent is comparable or indistinguishable to a near-optimal algorithm tuned to play this game and is indistinguishable to a human player. A comparison with humans is particularly interesting because, one of the primary aims of designing autonomous agents is to assign tasks otherwise performed by humans with sufficient guarantee for equivalent or better performance. In that respect, it is an important hypothesis to determine the reliability of the algorithm. Hence we have the following hypothesis:

**Hypothesis 6.1.1** *The performance of a coherence-driven agent is indistinguishable or comparable to the performance of humans and algorithms tuned to play the game.*

## 6.2 Design of Players

In this section, we discuss the three types of players in the experiment. For the *coherence* and near-opt players, we discuss their respective design.

### *near-opt* Agent Player

A *near-opt* agent is powered with an algorithm specifically tuned to the experimental set-up. It makes a plug action only if the battery charge is below 0.4 and the current cell contains a plug. In all other cases a *near-opt* agent makes a move action. In case of a move action, a *near-opt* agent first tries to move to a cell that has an option to *plug*. It achieves this by leveraging a crucial information: *if there is no density change after the last move, the last action was a move action, and cell X before the last move contained a plug, then a good target cell for the next move is cell X*. If any of the conditions is not satisfied, then a *near-opt* agent makes a move to a random cell. In Table 6.2, we illustrate how a *near-opt* agent chooses between move or plug action for the current round.

| **Variable** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| density change | no | no | no | no | no | no | yes | yes |
| can plug current cell | no | no | yes | yes | yes | yes | no | yes |
| can plug previous cell | no | yes | no | no | yes | yes | - | - |
| current energy level | - | - | $\leq 0.4$ | $> 0.4$ | $\leq 0.4$ | $> 0.4$ | $\leq 0.4$ | $> 0.4$ |
| action | move random | move to previous | plug | move random | plug | move to previous | plug | move to random |

### *coherence* Agent Player

A *coherence* agent implements a coherence-based architecture (see Section 5.3). Hence, such an agent selects between actions to plug and move based on coherence maximisation on its coherence graph. An action is chosen that is in the accepted set and is more preferred (higher degree) than other actions. For example, a move action is chosen if it is in the accepted set while the action to plug

is in the rejected set or in the accepted set with a lower preference value. We illustrate some of the technical details as we go through the design of a coherence agent and its reasoning following the procedure illustrated in Section 5.3.

Based on the coherence framework, a *coherence* agent represents elements of its theory (domain knowledge that indicates how to get its desire satisfied) as coherence graphs. This domain knowledge is encoded in theories of a belief, desire and intention logic as described in Section 5.3. For example, $(B(move \rightarrow points), 1)$ is a belief that a move will fetch a point with a confidence 1. Some of the theory elements of the agent are as given below:

$$(Dpoints, 1.0)$$
$$(B(move \rightarrow points), 1.0)$$
$$(B(energy \rightarrow move), 1.0)$$
$$(Bplug, (1 - e_s * d_s)/2))$$
$$(B(plug \rightarrow energy), 1.0)$$
$$(Bmove, e_s)$$
$$(Imove, x) \wedge (Iplug, x) \rightarrow \bot$$

Since there is no formal mechanisms to generate values for atomic graded cognitions, we have tried to match reality as well as possible. The only cognitions we will discuss here are $(Bplug, (1 - e_s * d_s)/2))$, $(Bmove, e_s)$ and $(Imove, r) \wedge (Iplug, r) \rightarrow \bot$, since the rest are intuitively obvious. $(Bplug, (1 - e_s * d_s)/2))$ translates to the belief to plug as a function of the energy and density sensor values. The intuition here is that the belief to plug increases with either a reduction in energy or a reduction in density and vice versa. Whereas, $(Bmove, e_s)$ is belief to move as a function of energy. Note that, the action to move does not depend on the density of plugs but only on the remaining energy. As these functions indicate, here we have tried to model a risk taking agent. However, it is possible to choose other functions. Formula $(Imove, x) \wedge (Iplug, x) \rightarrow \bot$ represents the mutual exclusiveness of action *move* and action *plug*.

A *coherence* agent then combines coherence graphs representing these theory elements using composition functions derived from bridge rules (see Section 5.2.2). Bridge rule

$$b_1 = \frac{C_B : (B(p \rightarrow q), \alpha), C_D : (Dq, \beta)}{C_D : (Dp, \min(\alpha, \beta))}$$

generates a new desire $p$ given the desire of $q$ and a belief that $p$ facilitates $q$ with minimum of the degrees. Bridge rule

$$b_2 = \frac{C_B : (Bp, \alpha), C_D : (Dp, \beta)}{C_I : (Ip, \min(\alpha, \beta))}$$

generates a corresponding intention given a desire and a belief that the desire is achievable (realistic agent). Note that bridge rule $b_1$ is very similar and motivated from the well known practical syllogism, "If I want $q$ and $p$ realises $q$, then I should intend to do $p$".

There is one single persistent desire for the agent, which is to earn points. Combined with the domain knowledge that $(B(move \rightarrow points), 1)$ (a move will fetch a point with a confidence degree 1) and applying bridge rule $b_1$ a new desire to "move" is generated. Further, combining the desire to move and the belief that *having energy enables move, i.e.,* $(B(energy \rightarrow move), 1)$, a new desire to have "energy" is generated. A third desire to "plug" is generated using the bridge rule and the belief that *plugging gives energy, i.e,* $(B(plug \rightarrow energy), 1)$. Hence, a chain of desires and their coherence links is generated.

Applying bridge rule $b_2$, a *coherence* agent generates the intention to move, intention to have energy and intention to plug. Further, as in the case of desires using $b_3$, the agent has that *the belief* $(B(energy \rightarrow move), 1)$*,* $(Ienergy, x)$ *and* $(Imove, x)$ *are coherently related.* The same is true of $(Ienergy, x)$ and $(Iplug, x)$. Hence, a chain of intentions and their coherence links are generated. Applying bridge rules repetitively, a coherence graph of the agent as in Figure 6.2 is constructed for the parameters $e_s = 0.8, d_s = 0.45, p_s = 1$.

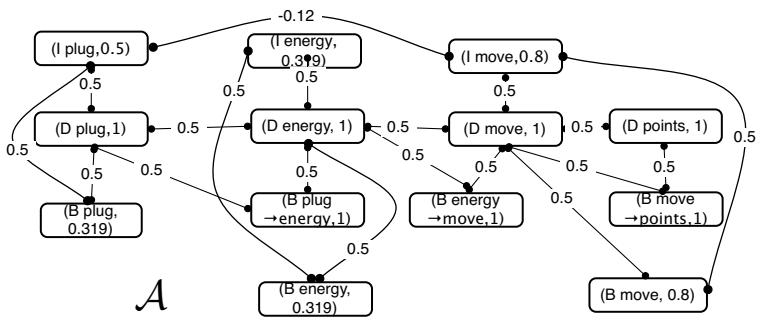


Figure 6.2: Coherence graph of agent $e_s = 0.8, d_s = 0.45, p_s = 1$

A *coherence* agent implements a deductive coherence function (see Definition 3.2.5) in prolog to generate a coherence graph from its theory presentations. Generating coherence maximising partitions from coherence graphs is implemented using a greedy algorithm combined with appropriate randomisation to avoid local optimas. Even though there are no performance indicators (see also a neural network approximation algorithm [Thagard, 2002]), it has given good results for testing purposes.

### *human* Player

Player type *human* is chosen from a random population. The experimental set-up and rules of the game are introduced to human players along with practice sessions prior to conducting the actual experiment. Due to the low skill level requirement, *human* players considered for this experiment may be at the level of experts. Moreover, to compensate for lack of concentration, the best score in three attempts under identical experimental conditions is taken.

## 6.3 Simulations

This section shows the simulation and results of the experiments that have been executed for the game scenario presented in Section 6.1.1. For the hypothesis presented, we explain the statistical analysis that has been realised on the experimental data and the results of such analysis.

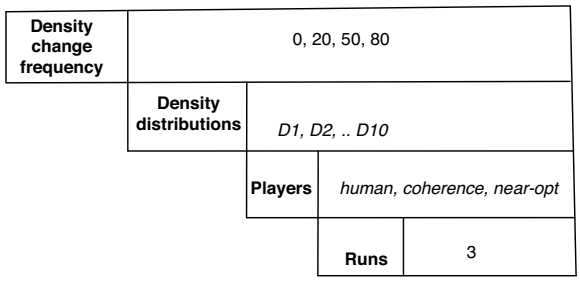


Figure 6.3: Simulation statistics

In order to test the hypothesis, we executed the following simulations. The game is played by a *coherence* agent player, 10 *human* players and a *near-opt* agent player. Simulations are run with the following parameters taking different values as shown in Figure 6.3. That is, for each density change frequency, the game is run 10 times (and for 10 different density and plug distributions) and in each of the runs, a *human*, a *coherence*, and a *near-opt* agent, each play 3 times. This is to compensate for possible human errors and the best performance value among the 3 repetitions is assigned as the performance of that player for the corresponding run. Before executing the statistical significance test we verified that the resulting data had a normal distribution through a Quantile-Quantile test, which is a precondition for certain statistical tests (Figure 6.4).

In order to test if there was a significant relationship between the independent variables, *i.e.,* the parameters in the simulation, and the dependent variable, *i.e.,* the *total score* of each player, we ran an analysis of variance (ANOVA) with the experimental data. The results of the ANOVA test demonstrate that all the independent variables were statistically significant with a p-value of 0.01 and 0.05. This information is indicative of the fact that the independent variables used in the experiment are significant to some extent, the parameter *density change frequency* more significant than the parameter *player type*. Though this is not in support with the hypothesis (with a border line significance), we verify further the variable relations using a Tukey test in order to analyse the variables in more detail.

In order to verify which variable values did better for each of the variables that interested us with respect to the hypothesis, we ran post-hoc comparisons using a Tukey test. It tests for differences in scale between two groups. The test is used to determine if one of two groups of data tends to have more widely

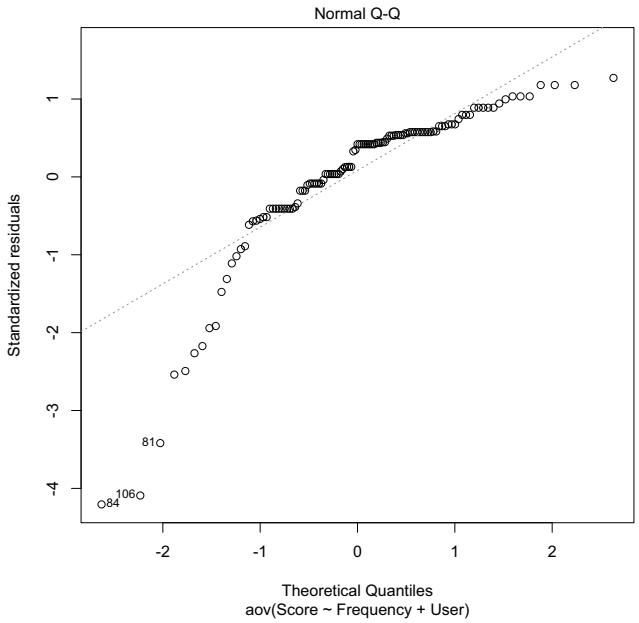


Figure 6.4: Result of Quantile-Quantile test showing the variables in a normal distribution

dispersed values than the other. In other words, the test determines whether one of the two groups tends to move, sometimes to the right, sometimes to the left, but away from the center (of the ordinal scale). The Tukey test for the parameter *player type* indicates that there is no significant difference between player type *human* (2) and player type *coherence agent* (1) since the mean difference lie close to 0.0 (see the results in Figure 6.5). However, the player type *near-opt agent* (0) does slightly better than both *human* players and *coherence* agent. This is expected, however, as the significance indicates, we cannot say there is a definite improvement. This is in support with our hypothesis and indicates that the performance of a coherence-driven agent and a human are indistinguishable whereas the performance of a coherence-driven agent and a near-opt agent does not vary greatly. The test on density change frequency as shown in Figure 6.6 indicates that there are no variables that are doing significantly better than another. Hence, even though ANOVA test shows there is a border line significance between player types, further analysis using Tukey test shows that the significance is due to the *near-opt* agents doing marginally better than the other player types. This is in support of our hypothesis.

With the help of the experimental evaluations, we have shown the feasibility of a coherence-driven agent and coherence-based decision making. To provide

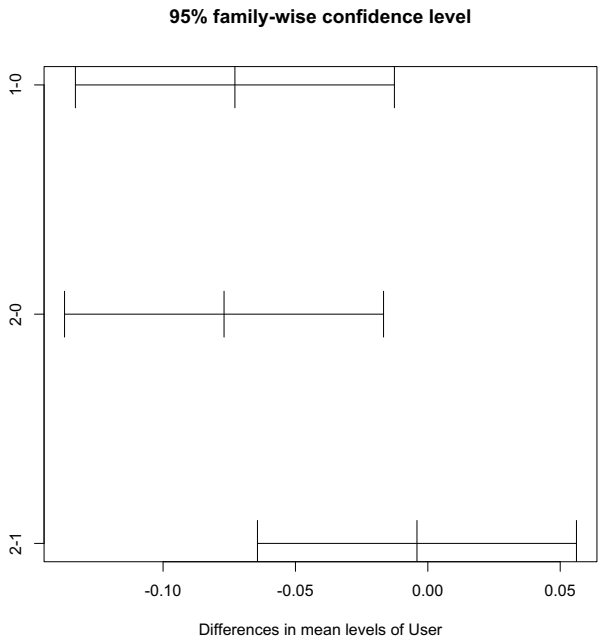


Figure 6.5: Tukey test to compare performance of player types *near-opt* (0), *coherence* (1), and *human* (2)

a better evaluation, it is necessary to make the game scenario sophisticated enough to have many action choices and conflicting cognitions so that near-optimal algorithms are hard to device.

## 6.4 Discussion

In this chapter, we have shown the feasibility of coherence-driven agents and have shown that their performances are comparable to that of humans and other specifically designed algorithms. This, we think is sufficiently significant in itself as the theory of coherence is known to be a motivational driver for humans. The computational formalism and framework verifies and reinforces the fact that the translation is sound and the theory can be used as the main motivational driver in artificial software agents. Further, humans are known to be adaptive to changing situations. The designed algorithm explicitly takes this into account. However, coherence-driven agents just as humans have adapted to changes in density levels without any adaptation tuning in the reasoning module. These indications tend to point to the potential of coherence-driven agents in modelling real world agents capable of taking autonomous decisions in the presence of

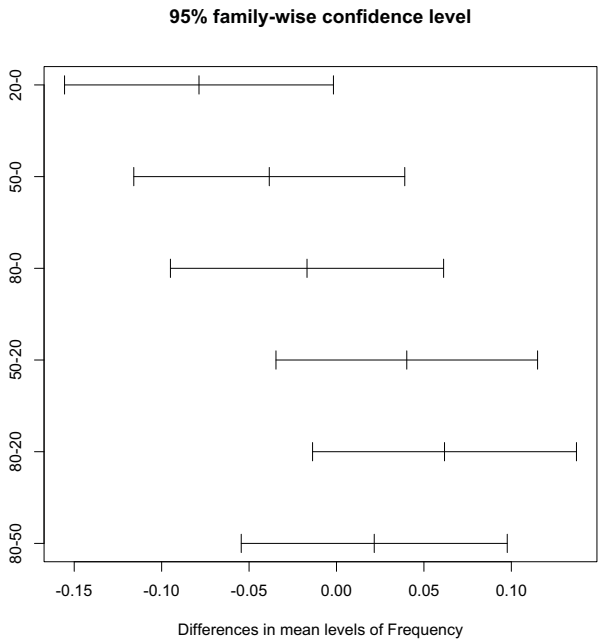


Figure 6.6: Tukey test to conduct a pairwise comparison of the performance under different density change frequencies $(0, 20, 50$ and $80)$.

conflicting interests.

In the following part of the book (Part III), we discuss the interactions of coherence-driven agents in a normative MAS, specifically those systems regulated by norms. For the purpose we extend the coherence-based architecture to incorporate normative considerations. We further define an argumentation framework to demonstrate that the social extension of a coherence-driven agent is intuitive, and has considerable advantage over agents built using argumentation systems following Dung's abstract argumentation system.

# Part III

# Extensions and Applications

# Chapter 7

# Autonomous Normative Agents

In this chapter, we define an *autonomous normative agent*. As discussed in Chapter 1, autonomous normative agents are autonomous agents situated in a normative environment such as a normative MAS and is capable of reasoning about norms autonomously. This includes a capability to reason about norm compliance as well as norm adoption. They are further capable of suggesting changes to existing norms or proposing new norms and reasoning about norm proposals of others. Hence, in this chapter, we address the research objective dealing with the construction of autonomous normative agents. Addressing this objective will prepare us to embark on normative agreements within a normative MAS, a topic that will be addressed in Chapter 8.

This objective is put in the context of an evolving normative system where autonomous normative agents propose, argue and deliberate over norms and change them over time based on consensus. The two primary tasks to assist this process is for agents to be able to propose norms and reason over norm proposals of other participants. Both tasks call for an extension of the coherence-based agent architecture so that agents are able to not only reason over cognitions but cognitions and norms put together. The generic coherence framework also needs to be extended with notions of *support and conflict sets*, which will aid coherence-driven agents to justify their reasoning.

We use a running example to illustrate the conceptual ideas discussed in the current chapter and in Chapter 8. We first introduce the example in Section 7.1, and then discuss the extension of the agent framework by introducing the *norm context* in the multi-context agent architecture in Section 7.2. In Section 7.3, we discuss the extension to the coherence framework and in Section 7.4 discuss our proposal on autonomous normative agents. We conclude with discussions in Section 7.5.

## 7.1 Example — Norm Deliberation

**Example 7.1.1** We choose the example in such a way as to illustrate the concepts on autonomous normative agents and multiagent norm deliberation (Chapter 8). The example is aimed at setting up a normative MAS for a discussion forum. We model two of the agents: $a$ and $b$ (both coherence-driven) forming an organising committee to discuss certain norms for regulating the discussion forum, especially on how often the participants may reply to each other's contributions. There are a set of social goals each agent is concerned with. This set may not be the same for both agents. However, a subset of these social goals are common knowledge to both agent which are referred to as *focal goals* and this set of goals is the reason why two agents think they should form a normative society. This set cannot be empty, otherwise there is no reason why the agents should come together. Here, we list the set of social goals (this includes all goals that any one of the agents has) and the list of focal goals:

- $e$ = efficiency (the discussion should not take too long)
- $c$ = coverage (the discussion should cover as much relevant material as possible)
- $f$ = fairness (the participants should be treated fairly compared to each other)
- $q$ = quality of contributions (the participants should be stimulated to write high-quality contributions).
- $x$ = A particular $Mr.x$ should not become a moderator (this is a private goal of agent $b$). Note that in this context a fair discussion of norms should not include any consideration to satisfaction of private goals. However, since a private goal is not revealed to anyone, its influence cannot be regulated as long as it doesn't enter any discussion.
- the only focal goal for the agents in this example is to achieve *efficiency* denoted as $e$.

Further, there are two norm proposals (methods for regulating the discussion forum) as given below:

- $\alpha$ — *make everyone reply as long as allowed by the moderator*
- $\beta$ — *enforcing that everyone gets one reply*

Initially, agent $a$ has the social goals $f, e$ and $c$. Further, agent $a$ believes that $\alpha$ promotes certain social goals. That is, $\alpha$ promotes efficiency ($\alpha \rightarrow e$) since a moderator can be trusted to keep discussions short. It also promotes fairness ($\alpha \rightarrow f$) since a moderator can be trusted to give experts more replies than novices. It has no particular effect on coverage or quality of contributions (since judging whether everything has been covered is too difficult for a moderator).

Agent $a$ however believes that, $\beta$ has an adverse effect on its social goal on coverage ($\beta \rightarrow \neg c$).

An initial theory of agent $a$, with the above information is as in Table 7.1. Note that the notation followed is as introduced in Section 5.2.1. Here the grades on beliefs represent a confidence degree on the corresponding belief. For example, $(B(\alpha \rightarrow e), 1)$ represents a belief that the proposal of $\alpha$ realises social goal $e$ with a confidence value 1). The degrees on desires are based on a priority ordering of desires. For example, for agent $a$, social goals $e, q$, and $c$ has a priority ordering $e > q > c$. Representation of norms follows the same pattern as that of other cognitive elements and will be introduced in Section 7.2. Degrees on norms also represent a priority ordering.

| *Theory* | $\mathcal{A}$ | $V \setminus \mathcal{A}$ |
|---|---|---|
| $\mathcal{T}_N$ | | |
| $\mathcal{T}_B$ | $(B(\alpha \rightarrow e), 1), (B(\alpha \rightarrow f), 0.9)$ $(B\beta \rightarrow \neg c, 0.8)$ | |
| $\mathcal{T}_D$ | $(De, 1), (Dq, 0.9), (Dc, 0.8)$ | |
| $\mathcal{T}_B{}^{\bullet}$ | $(B(\alpha, 1), (B(e, 1), (Bf, 0.9)$ $(B\beta, 1), (B\neg c, 0.8), (Bc, 0.8)$ | |

Table 7.1: The initial theory of Agent $a$

Agent $b$ however, believes that by enforcing $\beta$, certain social goals can be achieved. In particular, $\beta$ promotes efficiency ($\beta \rightarrow e$) and quality of individual contributions ( $\beta \rightarrow q$). The reason why this promotes quality of contributions is that with just one possible reply everyone will make it as good as possible, since they will not get a second chance. It has no net effect on fairness since on the one hand everyone gets the same number of replies (which is fair) but on the other hand an expert in the field will get less opportunity to say what he wants to say than a layman (which is unfair). Finally, $b$ believes that enforcing $\alpha$ may hinder achieving its private goal $x$. Put together, agent $b$ has the initial theory as in Table 7.2.

| *Theory* | $\mathcal{A}$ | $V \setminus \mathcal{A}$ |
|---|---|---|
| $\mathcal{T}_N$ | | |
| $\mathcal{T}_B$ | $(B(\beta \rightarrow e), 1), (B(\beta \rightarrow q), 0.9)$ $(B(\alpha \rightarrow \neg x), 0.8)$ | |
| $\mathcal{T}_D$ | $(De, 1), (Dq, 0.9), (Dx, 0.8)$ | |
| $\mathcal{T}_B{}^{\bullet}$ | $(B\beta, 1), (Be, 1), (Bq, 0.9)$ $(B\alpha, 1), (B\neg x, 0.8), (Bx, 0.8)$ | |

Table 7.2: The initial theory of Agent $b$

Given this background, we show how agents $a$ and $b$ may generate norm proposals and evaluate the proposals of the other for the purpose of regulating the discussion forum.

## 7.2 Norms

Normative behaviour in a normative MAS is generally described by using deontic constraints, such as obligations, permissions and prohibitions. Just as we have graded cognitions for an agent, the norms we consider also are graded. Grades in general add more richness to the semantics, and, in particular in the case of norms, grades help to understand the relative importance of a norm within a system of norms. A graded norm is interpreted in terms of its priority, measured in terms of the value it generates in a normative multiagent system. This value can be determined by the social goals it helps in achieving. In this book, we have assumed such a measure, however there are formalisms which detail how this could be done. A particularly interesting formalism in this context is that based on Atkinson in which the author proposes an action selection framework based on argumentation [Atkinson, 2005a]. The priority among different and possibly conflicting arguments is based on the social values promoted or demoted by action-options supported by these arguments. According to this formalism, each scenario may have different sets of values and depending on the value ordering, one action may be preferred to another. A preference ordering over actions can thus be generated based on a set of value orderings promoted by the enactment of the actions. This technique can be borrowed to generate preference ordering of norms with a corresponding mapping of actions to norms.

Another interpretation of grades in a norm is in terms of a probability of compliance. This probability can be used to prioritise norms, which helps an agent to understand more prevailing norms over others and those norms that may not be in use. This interpretation may be linked with the earlier interpretation by the reasoning that norms that are complied more often have more priority. There are however logics to treat both these interpretations. We in this book use the value-based preference ordering of the norms since later on in Chapter 8, we define a MAS that decides on a set of norms on the basis of their usefulness in achieving social goals. However, for a different purpose it may be more suitable to use the norm compliance interpretation of grades. It is also possible to combine the two interpretations.

### Norm Context

Similar to the cognitive contexts discussed in Chapter 5, a norm context $\mathcal{K}_N = \langle L_N, A_N \vdash_N \rangle$ consists of a norm language $L_N$, a set of axioms $A_N$ and an MDR $\vdash_N$ defining the logical system, together with a theory presentation $\mathcal{T}_N \subseteq L_N$ of the context. Since we have graded norms, we use a particular many valued extension of deontic logic as the norm logic. We base the norm logic on the work of Godo et al. [Dellunde and Godo, 2008] on necessity-valued deontic logic[1]. Further, the norm theory $\mathcal{T}_N$ gives rise to a coherence graph whose nodes are graded formulas of the norm language. We call this graph a *norm*

[1]The necessity-valued extension allows to attach preference or priority degrees to norms, and because we deal with preferences and priorities rather than probability, we use this particular extension here and not the probability-valued counterpart.

*graph* which is realised by extending the deductive coherence function. In the following subsection, we discuss the norm logic and the norm graph.

**Norm Logic**

In order to define a *norm graph*, we need to first define a norm logic $\mathcal{K}_N = \langle L_N, A_N, \vdash_N \rangle$. As mentioned previously, we define $\mathcal{K}_N$ as a graded deontic logic to represent and reason with norms. We define the norm language $L_N$ by extending a classical propositional language $L$ defined upon a countable set of propositional variables and connectives $\neg$ and $\rightarrow$. In particular, $L_N$ is defined as a fuzzy modal language over Standard Deontic Logic (SDL) to reason about the necessity degree of deontic propositions. The language, axioms and deduction relation are defined similarly as in the case of the belief logic. As in the case of the belief logic, given $\varphi$ is a SDL formula, $L_N$ contains formulas of the form $\bar{r} \rightarrow_L N\varphi$ where $r \in [0, 1]$. A formula $\bar{r} \rightarrow_L N\varphi$ expresses that the preference degree of norm $\varphi$ is at least $r$. We shall use the notation $(N\varphi, r)$ for this kind of formulas, and call them *graded norms*. $L_N^* \subseteq L_N$ denotes the set of all graded norms of $L_N$. Furthermore, we shall only consider theory presentations $\mathcal{T}_N \subseteq L_N^*$ expressed using graded norms.

Some examples of formulas in a graded normative language $L_N^*$ are given below. Also, to keep uniformity with the belief, desire, and intention languages as described above, we adopt a slightly different notation from that given in [Dellunde and Godo, 2008]:

$$\big(O(uses(john, public\_transport) \rightarrow validates(john, ticket)), 0.8\big)$$

means that, the priority is (at least) 0.8 that it is obligatory that, if John uses public transport, John validates the ticket;

$$\big(O(is\_citizen\_of(anna, utopia) \rightarrow pays\_taxes(anna, utopia)), 1\big)$$

means that, the priority is (at least) 1 that, it is obligatory that, if Anna is a citizen of Utopia, Anna pays taxes to Utopia.

We extend the definition of the coherence-driven agent to define a coherence-driven normative agent as below.

**Definition 7.2.1** *A* coherence-driven normative agent *a is a tuple* $\langle \{C_i\}_{i=\mathrm{B,D,I,N}}, B, \mathsf{cohgraph}, \mathsf{compfun} \rangle$ *where* $\{C_i\}_{i=\mathrm{B,D,I,N}}$ *is a family of contexts,* $B \subseteq \mathcal{B}$ *is a set of bridge rules,* $\mathsf{cohgraph} : \{C_i\}_{i=\mathrm{B,D,I,N}} \rightarrow \mathcal{G}$ *maps contexts to coherence graphs, and* $\mathsf{compfun} : 2^B \rightarrow \mathcal{G}^{(\mathcal{G}^4)}$ *maps sets of bridge rules to composition functions that take a quadruple of graphs to a graph.*

## 7.3 Extending the Coherence Framework

One of the desired characteristics of any reasoning or decision making system is its ability to explain or justify reasons or decisions. This is especially relevant

when the decision or reasoning needs to be explained to an external entity. In the context of collective decision making, for example, for reaching norm consensus in this case, it is important that agents are able to produce justification or support for their norm proposals. Later, to evaluate a norm proposal, an agent may need to know the arguments supporting the proposal. While the need for transparency is clear, one of the criticisms raised against coherence-based decision making is the lack of justification behind a decision. To counter this criticism we introduce two simple notions *a support set* and *a conflict set* of a node.

The support is defined in terms of the coherence links between nodes. The intuition is that if two nodes have a positive coherence between them, then they reinforce or give support to each other. Therefore, the support set of a node is all nodes that have a positive coherence link with the node concerned. However, the support set is always calculated with reference to a partition. This is due to the fact that the elements of the rejected set cannot be part of a support set. This is analogous to the concept of a justification in an argumentation system [Dung, 1995]. If an argument (node) is justified by a defeated argument (nodes from the rejected set), then the first argument itself does not hold as the justification is already defeated. Hence, we take only those nodes that are part of the accepted partition.

Note however that, there is a conceptual difference between coherence-based support and justification-based support. A justification-based support states deductive implications leading to a claim, whereas a coherence-based support cannot always construct such a chain of implications. Nevertheless, it is valid, since coherence of individual edges arise from valid implications. Another reason why a coherence-based support should be taken seriously is that one of the properties of a justified collection of arguments is its coherence.

**Definition 7.3.1** *Given a coherence graph $g = \langle V, E, \zeta \rangle$ and given a coherence maximising partition $(\mathcal{A}, V \setminus \mathcal{A})$,* the support set of a node $v$ *is given by*

$$S(v) = \{w \in \mathcal{A} \mid \zeta(\{v, w\}) > 0\} \tag{7.1}$$

*The support set of a set of nodes $W$ is defined as $S(W) = \bigcup_{w \in W} S(w)$.*

The support set of node $(B\alpha, 1)$ in Example 7.1.1 is the set $\{(B(\alpha \rightarrow e), 1), (B(\alpha \rightarrow f), 0.9), (Be, 1), (Bf, 0.9)\}$ as shown highlighted in Figure 7.1. Note that all nodes are part of the accepted set for the partition shown in the graph.

Similar to providing justification, it may also be necessary to provide reasons why a particular decision or reasoning is countered or attacked. As in the case of a support set, if two nodes have a negative coherence between them, then they oppose each other. However, it is not enough to simply collect those nodes that are in conflict with the node in question, but also those that provide support for the conflict nodes. Hence, the conflict set of a node consists of the set of nodes that is in conflict with the concerned node together with its support set. This

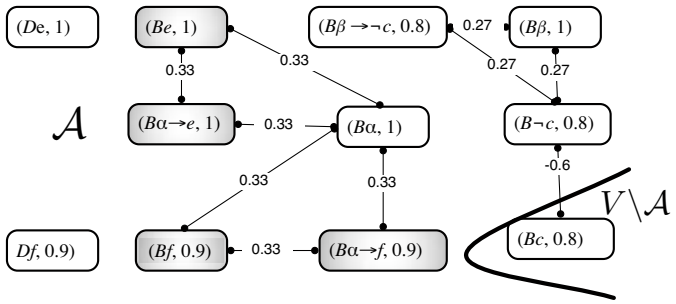


Figure 7.1: The support set of node $(B\alpha, 1)$ of agent $a$

is similar to the attack relation in argumentation systems in which to defeat an argument, the attacking argument should not itself be defeated.

**Definition 7.3.2** *Given a coherence graph $g = \langle V, E, \zeta \rangle$ and given a coherence maximising partition $(\mathcal{A}, V \setminus \mathcal{A})$,* the conflict set of a node $v$ *is given by*

$$C(v) = S(\{w \in V \mid \zeta(\{v, w\}) < 0\}) \tag{7.2}$$

The conflict set of node $(Bc, 1)$ in Example 7.1.1 is the set $\{(B\neg c, 0.8), (B(\beta \rightarrow \neg c), 0.8), (B\beta, 1)\}$ as shown highlighted in Figure 7.2.

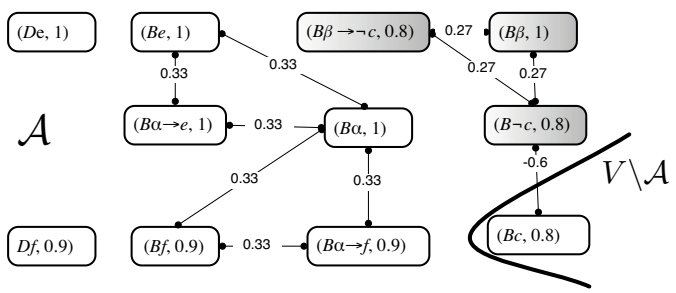


Figure 7.2: The conflict set of node $(Bc, 0.8)$ of agent $a$

Once an agent generates candidate norms and selects a norm to be proposed, the support set of the proposed norm helps to justify the proposal. Similarly, in the case of norm evaluations, the support set helps to evaluate an incoming norm proposal, while if rejected in its own coherence graph, a conflict set helps to counter the proposal.

## 7.4 Reasoning about Norms

In this section, we elaborate how an autonomous normative agent can generate new norm proposals and evaluate norm proposals made by others in a normative MAS. Normative MAS considered here are those that exist to realise a set of social goals. In this context, we make a distinction between social and private goals from the perspective of an agent. Social goals are what an agent thinks is good for its society while private goals are what the agent thinks is good for itself. Further, in a normative MAS, these social goals are realised through a set of norms. Even though we make no assumptions about the motivations of an agent, norm proposals should be justified as means to realise one or more social goals. Here we discuss norm proposals and evaluations of them in general without any reference to existing norms. This may be applied to situations where an agent has to generate a completely new norm or make modifications to an existing norm. The internal deliberation both in the case of new norm proposals and evaluation of norm proposals of others is guided by the process of coherence maximisation.

### 7.4.1 Norm Generation

Here we give a general account of the process of norm generation by a coherence-driven agent. We specify conditions under which an action is obliged, prohibited, and pairs of them mutually excluded. Conte et al. specify certain conditions under which an agent adopts a norm [Conte et al., 1999]. Among other things, it has to satisfy the condition that the norm will be instrumental to solving some of the social or private goals of the agent. We extend this principle to specify conditions under which a new norm is generated. A new norm, we claim, stems from an unsatisfied social goal and a belief that certain actions under certain conditions (or none) can achieve this goal. We exclude the case where a norm is instrumental to satisfying only a private goal, since this will generate too many norms that are not likely to be accepted by other participants. Hence we stick to those norms that are instrumental to satisfying one or more social goals.

In the example 7.1.1, agent $a$ believes that $\alpha$ helps to realise social goal $e$ $((B\alpha \rightarrow e, 1))$ and desires $((De, 1))$ to be realised. Given this theory, norm $\alpha$ is inferred in the norm context $((O\alpha, 1))$.

In general we have the bridge rule that says *if the goal context implies a social goal $\psi$ and the belief context implies a belief $\phi \rightarrow \psi$ then the normative context infers an obligation $\phi$.*

**Rule 1:**
$$\frac{C_B : (B(\phi \rightarrow \psi), r), C_D : (D\psi, s)}{C_O : (O\phi, f(r, s))}$$

(For the examples used in this chapter, $f(r, s)$ is $\min(r, s)$.) If applied naively, this bridge rule will result in too many obligations: if there is more than one way to achieve $\psi$, then all of them will be turned into obligations, which would over-constrain the normative system: what we want instead is to make only one

way to achieve the social goal obligatory, in order to increase the agent's degree of autonomy. Another aspect not taken into account by this bridge rule is that, realising $\phi$ may frustrate another social goal, i.e., it may hold that $\phi \rightarrow \neg\psi'$ where $\psi'$ is another social goal of the agent.

To deal with these problems, the obvious similarity can be exploited between this bridge rule and the well-known practical syllogism "If I want $\psi$, and $\phi$ realises $\psi$, then I should intend to do $\phi$". Walton formulated this as one of his presumptive argument schemes, with as main critical questions "are there other ways to realise $\psi$?" and "does $\phi$ also have unwanted consequences?" [Walton, 1996]. In recent years several AI researchers have formalised this argument scheme in formal argumentation systems (e.g. [Atkinson, 2005b, Bench-Capon and Prakken, 2006, Amgoud and Prade, 2009]). The key idea here is that positive answers to Walton's two critical questions give rise to counterarguments.

Our task is to model the same idea in the coherence approach. Since coherence theory is developed to make sense of such contradictions between pieces of information and identify those that cohere maximally, modelling the above scenario is natural using this theory. Coherence maximisation partitions the cognitive elements including the obligations in such a way that, the most coherent set of cognitions and obligations is selected. Note that the basic relationship we model here is that between goals and norms, in which different ways to achieve the same goal negatively cohere with each other. However, the present coherence framework uses only deduction as the underlying relation, in which the set $\{\phi \rightarrow \psi, \phi' \rightarrow \psi, \phi, \phi'\}$ is consistent (here $\phi$ and $\phi'$ are different ways to achieve goal $\psi$). Hence, we add an additional constraint to make these alternatives inconsistent. That is, for each goal $\psi$ in an agent's desire context, we consider the set of all implications $\phi_1 \rightarrow \psi, \cdots, \phi_n \rightarrow \psi$ in its belief context and we add formulas $\neg(O\phi_i \& O\phi_j)$ to its norm context for all $\phi_i$ and $\phi_j$ such that $1 \leq i < j \leq n$. Then, two obligations $O\phi_i$ and $O\phi_j$ negatively cohere with each other since they are alternatives[2]. This kind of constructs have also been used in argumentation-based negotiations earlier (e.g. see [Parsons et al., 1998]).

**Rule 2:** $$\frac{C_B : (B(\phi \rightarrow \psi), r), C_B : (B(\phi' \rightarrow \psi), r'), C_D : (D\psi, s)}{C_O : (\neg(O\phi \wedge O\phi'), f(r, r', s))}$$

Note that $f(r, r', s)$ is calculated as in Definition 5.2.3.

This method deals with the first of Walton's critical questions of the practical syllogism (are there alternative ways to realise the same goal?). To deal with his second critical question (does $\phi$ also have unwanted consequences?) a bridge rule is needed that expresses the negative version of the practical syllogism: if the goal context implies a social goal $\psi'$ and the belief context implies a belief $\phi \rightarrow \neg\psi'$ then the normative context contains an obligation $\neg\phi$.

[2]Explanatory coherence includes such a principle that states "if $\phi$ and $\phi'$ both explain a proposition, and if $\phi$ and $\phi'$ are not explanatorily connected, then $\phi$ and $\phi'$ are incoherent with each other ($\phi$ and $\phi'$ are explanatorily connected if one explains the other or is together they are used to explain some other proposition)."((Principle $E6$), [Thagard, 2002]

| **Rule 3:** | $\dfrac{C_B : (B(\phi \rightarrow \neg\psi), r), C_D : (D\psi, s)}{C_O : (O\neg\phi, f(r,s))}$ |
|---|---|

(For the examples used in this chapter, $f(r,s)$ is $\min(r,s)$.) Then, in cases where an action achieves some but frustrates other social goals, our deductive coherence measure makes the obligations that result from the positive and negative version of the practical syllogism negatively cohere with each other.

In the Example 7.1.1, agent $a$ has the belief that enforcing $\beta$ hinders realising goal $c$ $((B\beta \rightarrow \neg c, 0.8))$, however desires to realise goal $c$ $((Dc, 0.8))$.That is, we have from Rule 3,

$$\frac{C_B : (B(\beta \rightarrow \neg c), 0.8), C_D : (Dc, 0.8)}{C_O : (O\neg\beta, 0.8)}$$

and hence the norm context is updated with a new obligation $(O\neg\beta, 0.8)$ as $\mathcal{T}_N = \{(O\alpha, 1), (O\neg\beta, 0.8)\}$. As a consequence, from Rule 2 we have that $(O\alpha, 1)$ and $(O\neg\beta, 0.8)$ negatively cohere and $\zeta(\{(O\neg\beta, 0.8), (O\alpha, 1)\}) = 0.8$ from Definition 5.2.3. Hence the updated theory of agent $a$ is as shown in Table 7.3 and the coherence graph with the new norms and corresponding links is as in Figure 7.3.

| *Theory* | $\mathcal{A}$ | $V \setminus \mathcal{A}$ |
|---|---|---|
| $\mathcal{T}_N$ | $(O\alpha, 1), (O\neg\beta, 0.8)$ | |
| $\mathcal{T}_B^{\bullet}$ | $(B(\alpha, 1), (B(e, 1), (Bf, 0.9)$<br>$(B\beta, 1),\ (B\neg c, 0.8)$<br>$(B(\alpha \rightarrow e), 1), (B(\alpha \rightarrow f), 0.9)$<br>$(B\beta \rightarrow \neg c, 0.8)$ | $(Bc, 0.8)$ |
| $\mathcal{T}_D$ | $(De, 1), (Df, 0.9), (Dc, 0.8)$ | |

Table 7.3: Updated theory of $a$ with deduced norm $(O\alpha, 1)$

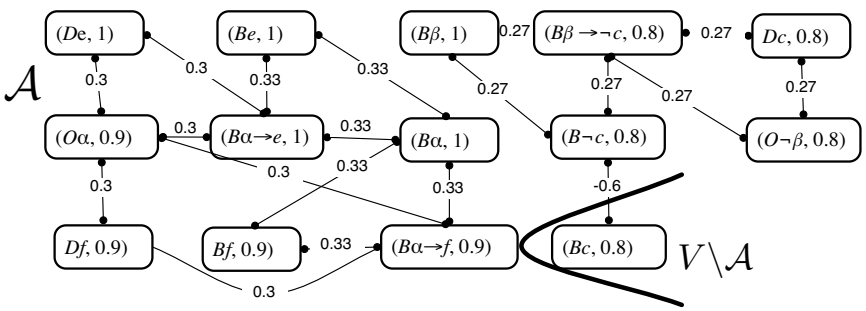


Figure 7.3: The coherence graph of agent $a$ with updated norm information

### 7.4.2 Generating a Norm Proposal

The application of bridge rules described in the previous section generates a set of possible norms. A new composition graph is generated from the coherence graphs updated with the set of possible norms and their closures as described in Section 5.2. The accepted set resulting from the coherence maximising partition of this graph is the base for chosing a norm among the generated norms. A norm proposal is then defined as a pair of the chosen norm and its support set. A coherence-driven agent with the multi-context architecture executes the following steps to generate a norm proposal:

1. it adds the generated norms to the theory $\mathcal{T}_N$ of the context $C_{\mathrm{N}}$;
2. it computes the deductive closure of ${\mathcal{T}^{\bullet}}_N$;
3. it expresses the contexts with newly closed theories as coherence graphs and computes the tuple $\bar{g} = \langle \mathsf{cohgraph}(C_{\mathrm{B}}), \mathsf{cohgraph}(C_{\mathrm{D}}), \mathsf{cohgraph}(C_{\mathrm{I}}), \mathsf{cohgraph}(C_{\mathrm{N}}) \rangle$ associated to them;
4. it computes the composite graph $g = \mathsf{compfun}(\bar{g}) = \iota_B(S^*(\bar{g}))$, where $S = \{\varepsilon_b \mid b \in B\}$;
5. it computes all coherence-maximising partitions $(\mathcal{A}_i, V \setminus \mathcal{A}_i)$, where $V$ is the set of nodes of the composite graph $g$;
6. it selects a coherence-maximising partition $(\mathcal{A}, V \setminus \mathcal{A})$ according to the criteria in Section 3.1;
7. it selects a norm $\rho = (O\alpha, r)$ such that $r = \max(s \mid (O\varphi, s) \in \mathcal{A})$;
8. it then generates the support set $S(\rho) = S((O\alpha, r))$.
9. it returns $(\rho, S(\rho))$ as the new norm proposal;
10. it returns null If $\rho$ is empty;

Highlighted in Figure 7.4 is the norm proposal of agent $a$ (norm $(O\alpha, 0.9)$ and its support set $\{(B\alpha \rightarrow e, 1), (De, 1), (B\alpha \rightarrow f, 0.9), (Df, 0.9)\}$). Note that due to the lower preference, norm $(O\neg\beta, 0.8)$ is not selected for proposal.

### 7.4.3 Evaluating a Norm Proposal

Evaluating an incoming norm proposal is again based on coherence maximisation. The agent introduces the received norm along with its support set into its respective coherence graphs and recalculates the composite coherence graph. Upon maximising coherence, if the norm falls in the accepted set of its coherence maximising partition, it accepts the norm proposal. Else it generates the conflict set of the norm taken from its coherence graph as the reason for rejecting the proposal. Given the norm proposal $(\rho, S(\rho))$ where $\rho = (O\varphi, r)$, and $S(\rho)$ a support set for $\rho$, a coherence-driven agent evaluates the proposal as follows:

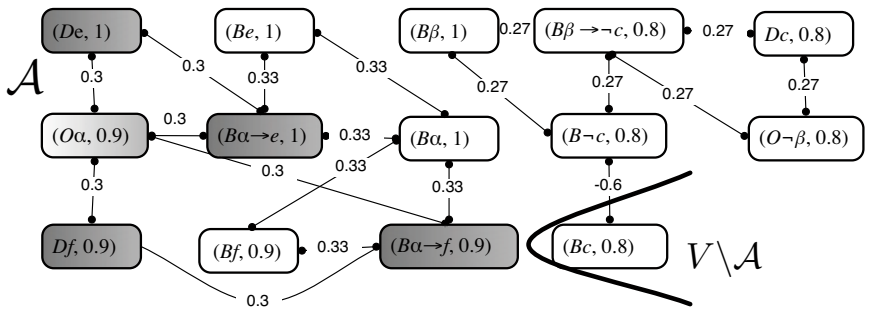


Figure 7.4: Norm proposal of agent $a$

1. it adds the received norm $\rho$ and elements of $S(\rho)$ to the theory $\mathcal{T}$ of the corresponding contexts $C_{\mathrm{B}}$, $C_{\mathrm{D}}$, $C_{\mathrm{I}}$ or $C_{\mathrm{N}}$;

2. it computes the deductive closure $\mathcal{T}^{\bullet}$;

3. it expresses the contexts with newly closed theories as coherence graphs and computes the tuple $\bar{g} = \langle \mathsf{cohgraph}(C_{\mathrm{B}}), \mathsf{cohgraph}(C_{\mathrm{D}}), \mathsf{cohgraph}(C_{\mathrm{I}}), \mathsf{cohgraph}(C_{\mathrm{N}}) \rangle$ associated to them;

4. it computes the composite graph $g = \mathsf{compfun}(\bar{g}) = \iota_B(S^*(\bar{g}))$, where $S = \{\varepsilon_b \mid b \in B\}$;

5. it computes all coherence-maximising partitions $(\mathcal{A}_i, V \setminus \mathcal{A}_i)$, where $V$ is the set of nodes of the composite graph $g$;

6. it selects a coherence-maximising partition $(\mathcal{A}, V \setminus \mathcal{A})$ according to the criteria in Section 3.1;

7. it accepts the norm proposal if $\rho \in \mathcal{A}$;

8. it rejects the norm proposal if $\rho \in V \setminus \mathcal{A}$ and generates the conflict set $C(\rho)$.

For the Example 7.1.1, we now show the evaluation of the norm proposal of agent $a$ by agent $b$. Table 7.4 shows the updated theory of agent $b$ with new norms according to the rules for norm generation, and Figure 7.5 shows the corresponding coherence graph.

Agent $b$ now receives the proposal of a norm $(O\alpha, 0.9)$ along with the support set $\{(B\alpha \rightarrow e, 1), (De, 1), (B\alpha \rightarrow f, 0.9), (Df, 0.9)\}$. $b$ incorporates the norm proposal into its coherence graph and re-evaluates its coherence. However, the coherence maximising partition classifies the proposed norm into the rejected set as in Figure 7.6. Hence, agent $b$ rejects the norm proposal and provides the reasons for the rejection as the conflict set as highlighted in the figure.

| *Theory* | $\mathcal{A}$ | $V \setminus \mathcal{A}$ |
|---|---|---|
| $\mathcal{T}_N$ | $(O\beta, 0.9), (O\neg\alpha, 0.8)$ | |
| $\mathcal{T}_B$ | $(B(\beta \rightarrow e), 1), (B(\beta \rightarrow q), 0.9)$ $(B(\alpha \rightarrow \neg x), 0.8)$ | |
| $\mathcal{T}_D$ | $(De, 1), (Dq, 0.9), (Dx, 0.8)$ | |
| $\mathcal{T}_B{}^{\bullet}$ | $(B\beta, 1), (Be, 1), (Bq, 0.9)$ $(B\alpha, 1), (B\neg x, 0.8)$ | $(Bx, 0.8)$ |

Table 7.4: The initial theory of Agent $b$

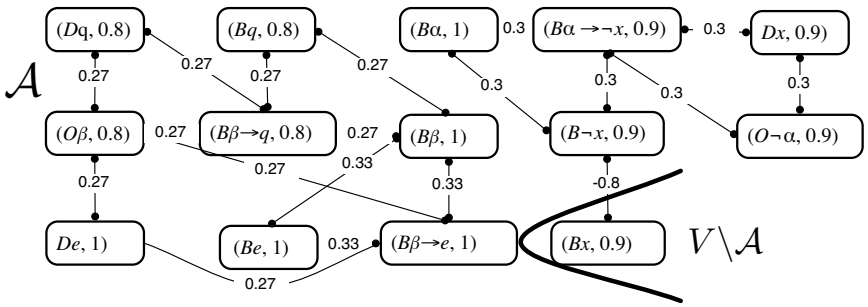


Figure 7.5: Initial Coherence graph of agent $b$ with the generated norms

It is interesting to note that the coherence maximising partition only rejects $(O\alpha, 0.9), (B\alpha \rightarrow f, 0.9)$ and $(Df, 0.9)$, the minimum number of elements to stay coherent.

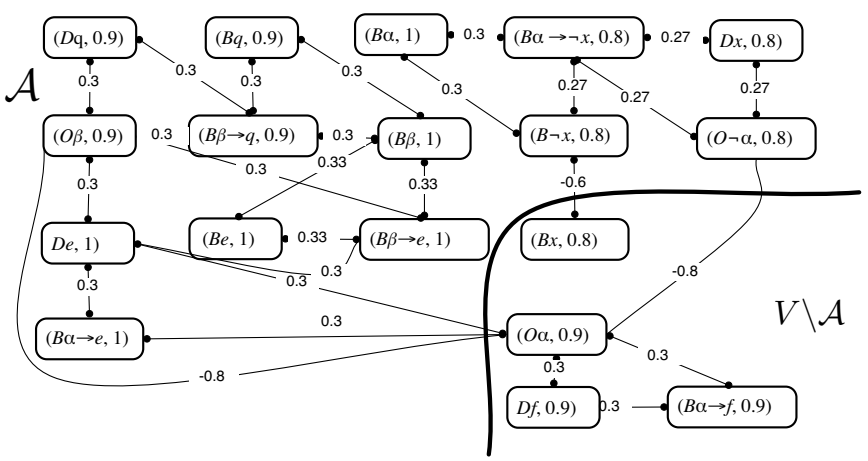


Figure 7.6: Evaluation of norm proposal of agent $a$ by agent $b$

## 7.5 Discussion

In this chapter, we have discussed the construction of autonomous normative agents. In particular, we have addressed generation and evaluation of norm proposals within a normative MAS. For the purpose, the coherence-based agent architecture was extended to include a normative context. Further, to aid the reasoning process and to overcome limitations on transparency, we extended the coherence framework to include notions of support and conflict sets. These aid the norm proposal and evaluation process by providing coherence-based justifications or reasons. These notions also help to evaluate a norm proposal, as an agent not only receives the norm being proposed, but also receives the support set(reasons) for the proposal.

In the following chapter, we explain the actual multiagent deliberation process, and define a dialogue protocol and conditions on convergence. We also distinguish the coherence-driven argumentation system from other argumentation systems for deliberation. The main point of differentiation is the introduction of a *joint coherence graph* on which computations can be performed. We detail these concepts and compare our approach with some of the leading research in the field.

# Chapter 8

# Multiagent Norm Deliberation

*"If you can find something everyone agrees on, it's wrong."*
Mo Udall

This chapter is a natural continuation of the previous chapter (Chapter 7) where we discussed normative reasoning within an autonomous normative agent, specially norm generation and evaluation. In this chapter, we further discuss the agent's interactions with a normative MAS. So far, in the literature, the interaction between these has been kept static, where a system designer would set up the norms governing the normative system, based on a set of social goals and impart to agents joining the normative system [Sierra et al., 2004]. As mentioned in Chapter 1, for the latest developments in areas such as the study of virtual organisations and communities, distributed environments like electronic institutions, MAS, and P2P networks, this static view of norms no longer suffices. This is due to the nature of such applications, which are essentially dynamic. Situational changes, changes in motivations (social goals), and hence changes in interactions that are acceptable vary over time. To regulate such societies, it is essential to capture the dynamics of norms, and to describe how they can be manipulated (i.e. revised, merged, institutionalised) [Boella et al., 2007, Boella et al., 2009]. Here, we try to understand the dynamic aspects of norms and the dynamic interaction between agents and the normative system. In particular, we are interested in normative MAS that can autonomously set up norms of regulation and propose and agree upon norm changes when necessary. As we have addressed the cognitive aspect (norm generation and evaluation) in Chapter 7, we address the social aspect of the norm, specifically how to reach consensus on a set of norms for regulation, which is the research question in this chapter:

> Given that each agent can propose and evaluate candidate norms, how can a group of agents reach consensus on a set of norms for self-regulation?

We continue to use Example 7.1.1 from Chapter 7 to illustrate how two coherence-driven agents reach agreement about regulating a certain aspect of their society. In this chapter, we propose a *dialogue system* to model the interaction among agents in a normative system. As mentioned in the introduction (Chapter 1), this dialogue system uses a coherence-driven argumentation to model the interaction among the agents. This differs from other typical argumentation systems in its basic notion of an argument. An argument in general consists of a *claim*, *a set of assumptions or premises* and *a method of reasoning or deduction* which relate the premises to the claim. In typical argumentation systems, the method of reasoning is deduction whereas here, the method of reasoning is coherence-driven. In deliberation dialogues, the participants usually jointly build a dialogue structure which reflect the history and the state of the dialogue. Here the joint structure is a *joint coherence graph* whose properties influence the progress of the dialogue. In Section 8.1, we discuss the languages, the joint coherence graph and the dialogue protocol which form the main components of the dialogue system proposed here. And in Section 8.2, we compare our dialogue system with some of the existing systems and summarise the conclusions.

## 8.1 Dialogue System

In this section we introduce the dialogue system to model coherence-driven argumentation for a group of agents to reach consensus on norms. When artificial entities take part in a conversation, the dialogue modeling is carried out with respect to a dialogue type they intent to use. Following the general outline of Hamblin [Hamblin, 1970] and Walton and Krabbe [Walton and Krabbe, 1995], four fundamental building blocks of any formal dialectical system can be identified:

1. the two participants, called the proponent and the respondent,
2. the types of moves (taking the form of various speech acts) that the two participants are allowed to make, as each takes his or her turn to speak,
3. the sequence of moves, in which the appropriateness of each move depends on the type of preceding move made by the other party,
4. the goal of the dialogue as a whole; the sequence of moves should ideally move towards the fulfillment of the goal as the dialogue proceeds.

Considering these building blocks for the dialectical system, we describe the three components of the dialogue system proposed here. A common shared language (based on item 2) is one of the basic assumptions in having a dialogue system in place. Agents also should share a topic language to carry out sensible deliberations. A second and essential component of the system is the dialogue protocol (addressing item 1, 3 and 4) which defines the interaction rules. Further, since the protocol depends heavily on a joint dialogue structure (joint coherence

graph), that represents the state of the dialogue and that acts as certain effect rules for directing the course of action in the dialogue, we have the joint coherence graph as the third component in the dialogue system. Hence, a dialogue system is as given in the following:

**Definition 8.1.1** *A dialogue system is a triple $\mathcal{D} = (L_c, Pr, \mathcal{J})$ where $L_c$, the communication language, is a set of locutions, $Pr$ is a protocol in $\mathcal{D}$, and $\mathcal{J}$ is the joint coherence graph representing the state of the dialogue.*

In this section we discuss the three components and illustrate them with the running example. We start by discussing a few assumptions we make in defining the dialogue system.

### 8.1.1 Assumptions

To make our discussion on the dialogue system concrete, we make a few assumptions about the objectives of the dialogue and the agent's awareness about these objectives. These are realistic assumptions likely to hold in most deliberation dialogues of the kind we are concerned here. However, these assumptions are made with the objective of setting up a normative system, and hence we do not discuss a norm proposal in the light of already existing norms. As discussed in the introduction, norm revision is another problem relevant to understanding the dynamics of norms. This objective is not taken into consideration in defining the dialogue system, although it may be incorporated later by enhancing the communication language $L_c$. While discussing the assumptions, we use the term *social goals* to refer to goals relevant for the society (for the group of agents) and *focal goals* to refer to mutually accepted social goals. All social goals may not be agreed upon by all the participants, but can become after deliberation. Focal goals are those social goals that are agreed upon a priori and the normative system is set up to realise these focal goals.

1. Dialogues are triggered by a set of mutually accepted social goals (the 'focal goals').

2. Dialogues are about how best to promote the achievement of these goals by enacting norms (in the hope that the agents of the relevant society will obey the norms and thus help realise the desired effects.)

3. During a dialogue additional social goals may be proposed by each agent and, if accepted by the other agent, norms for these additional goals may be proposed, or norm proposals for the focal goals may be evaluated in terms of their effect on the additional social goals.

4. Besides social goals, the agents may also have their own private goals. These are not made public during a dialogue but may be used internally by the agent that holds them to decide about making or accepting a proposal.

5. Existence of a shared communication language $L_c$ among the participants of the dialogue system and a shared topic language $L_t$ to express the contents of the dialogue.

### 8.1.2 Communication and Topic Languages

The first essential component of a dialogue system are languages for communication and for representation of the topic or content. We first discuss the topic language $L_t$ followed by the communication language $L_c$. Since our agents are coherence-driven as described in Part II, based on a multi-context architecture, the languages for expressing theories in each context are already defined. Hence we have a belief language $\mathcal{L}_B$, a desire language $\mathcal{L}_D$, a intention language $\mathcal{L}_I$ and a norm language $\mathcal{L}_N$ for expressing belief, desire, intention and norm theories respectively. Further, since the dialogues will only contain elements drawn from these theories, the topic language ideally is a union of the above languages. Hence we define $L_t$ as consisting of the union of the agents' context languages for beliefs, desires, intentions and norms as in the following definition.

**Definition 8.1.2** *Given a finite group of agents $\{a_1, a_2, \cdots, a_n\}$, each with a coherence-based normative architecture, and whose belief, desire, intention and norm theories expressed in belief language $\mathcal{L}_B$, desire language $\mathcal{L}_D$, intention language $\mathcal{L}_I$ and norm language $\mathcal{L}_N$ respectively, the topic language $L_t$ is given by the union of the context languages*

$$L_t = \mathcal{L}_B \cup \mathcal{L}_D \cup \mathcal{L}_I \cup \mathcal{L}_N$$

We now define $L_c$, the shared communication language, for a deliberation dialogue. In the special case of this dialogue system, the language $L_c$ is kept very simple. Since the dialogue system is designed keeping in mind a deliberation dialogue, the only element to be expressed in the language $L_c$ are the *arguments*. An argument (which will be formally defined below) consists of expressions $\rho$ *since* $\Gamma$ such that $\rho$ and all elements of $\Gamma$ are well-formed formulas of $L_t$. $L_c$ can be enriched with other locutions for a more complex dialogue as in [Prakken, 2005a].

**Definition 8.1.3** *A communication language $L_c$ is a set of* locutions *such that*

1. *for all $\Gamma(\neq \emptyset) \subseteq L_t$ and $\rho \in L_t$ we have $\rho$* since $\Gamma \in L_c$

For example

$$(B \textit{ car will not start}, 1) \textit{ since } (B \textit{ engine trouble}, 1)$$

represents a claim that *the belief that the car will not start* based on the support that *the belief that there is engine trouble* is a valid expression in $L_c$ since both $(B$ *car will not start*$, 1)$ and $(B$ *engine trouble*$, 1)$ belong to $\mathcal{L}_B$ and hence belong to $L_t$.

Even though any expression of the form $\rho$ *since* $\Gamma$ is valid in the language and is a valid argument, the protocol constraints the choices for $\Gamma$ and $\rho$ as will be evident later in the chapter. A dialogue move and a dialogue can be explained in terms of the elements of $L_c$. A *move* is a pair $(a, x)$ where $x$ is an expression from $L_c$ and $a$ is the agent who utters $x$ (sometimes we will abuse notation and refer to $x$ only as a move, leaving the speaker implicit). A *dialogue* is a sequence of moves. For any dialogue $d = m_1, \ldots, m_i, \ldots$ the sequence $m_1, \ldots, m_i$ is denoted by $d_i$, where $d_0$ denotes the empty dialogue. For any dialogue $d$ and move $m$ the notation $d, m$ stands for the result of appending $m$ to $d$, i.e., for $d$ as continued by $m$.

So far in Chapter 7 and Chapter 8, we have only introduced dialogue moves that contain norm proposals (dialogue moves of the form $\rho$ *since* $\Gamma$ where $\rho$ is exclusively a norm). However, dialogue moves can be about beliefs, desires or actions, and we refer to the special case where $\rho$ is a norm as a *norm proposal* and is defined as below: Since we refer to norm proposals exclusively, we define them formally here.

**Definition 8.1.4** *$\rho$ since $\Gamma$ is a* norm proposal *by agent $a$ in dialogue $d$ if $\rho \in \mathcal{L}_N$.*

Hence, the norm proposal of agent $a$ from Example 7.1.1 may be formally expressed as

$$m_1 = (O\alpha, 0.9) \textit{ since } \{(B\alpha \rightarrow e, 1), (De, 1), (B\alpha \rightarrow f, 0.9), (Df, 0.9)\}$$

Also when dialogue moves are norm proposals, they may be proposed to realise one or more social goals. Here we define when a norm proposal realises (addresses) a social goal. We use these special structures later in constructing the joint coherence graph and defining the protocol. We say that a norm proposal addresses a social goal when the dialogue move is for a norm proposal with the support set of the norm contains the referred social goal as one of the elements.

**Definition 8.1.5** *A social goal $(D\psi, t)$ is* addressed *by a norm proposal by an agent $a$ in $d$ if $d$ contains a move $(a, x)$ where $x$ is of the form $(O\phi, r)$ since $(B(\phi \rightarrow \psi), s), (D\psi, t)$ ( applying Rule 1 of Section 7.4).*

### 8.1.3 Joint Coherence Graph $\mathcal{J}$

A general feature of the dialogue system is that it is for 'theory building' dialogues where participants jointly build a common structure that generally stores everything that has been exchanged. This is because we use *deliberative dialogues* in our dialogue system. This is in accordance with the classification of dialogue types by Walton and Krabbe [Walton and Krabbe, 1995] according to which Participants of *Deliberation Dialogues* collaborate to decide what action or course of actions should be adopted in some situation. Here, participants share a responsibility to decide the course of action, or, at least, they share a willingness to discuss whether they have such a shared responsibility. Note that

the best course of action for a group may conflict with the preferences or intentions of each individual member of the group; moreover, no one participant may have all the information required to decide what is best for the group.

The reason for choosing a theory-building approach is that in deliberations about promoting the achievement of social goals by enacting norms, the public understanding of a problem is crucial: since the goals addressed are social and the norms bind everyone within the relevant society, the reasons for a consensus should ideally be public. This contrasts with argument-based negotiation [Rahwan et al., 2003a], where the negotiating parties are self-interested so that all that counts is whether an argument persuades the hearer to do something in the dialogue (like accepting or revising a proposal) that is beneficial to the speaker. In consequence, protocols for argument-based negotiation usually are not of the theory-building kind but define the outcome of a dialogue purely in terms of explicit acceptances and rejections. Some persuasion protocols are also of that kind, which is suitable when the participants' only goal is to win a dispute. However, when public interests are at stake, a theory-building approach arguably is better.

The idea of theory-building dialogues is not new (see e.g. [Gordon, 1994]), but in most current dialogue systems for argumentation the theory built during a dialogue is a set of arguments or some related structure (such as a dialectical graph). By contrast, in this framework, the theory built during a dialogue is a coherence graph which we call a *joint coherence graph*. The agents' arguments can contain norm proposals (refer Section 7.4) or can be about goals or matters of belief. The agent's arguments are then incorporated in the joint coherence graph in nodes corresponding to the conclusions and premises of the arguments, and in the relevant positive and negative constraints between these nodes. In fact, as you will see later, the protocol will require of arguments that when added to the joint coherence graph, there is indeed a positive coherence in the graph between the argument's premises and conclusion. Thus in our system, the notion of an argument is not basic but derived: the basic reasoning/inferential structure is not a set of arguments but a coherence graph, and the inferential machinery applied to the joint theory is not an argument-based logic but a coherence calculus.

The joint graph is then used by the protocol to define turntaking, relevance of moves and the dialogue outcome, in ways explained while discussing the protocol. Besides the dialogue's joint coherence graph each agent also has its own internal coherence graph (which may also be updated or revised during a dialogue but which remains hidden to the other agent). This graph is used by the agent to make its internal decisions about what to say in the dialogue (e.g. whether to make or accept a certain proposal). An agent's private goals are incorporated into its internal coherence graph. Note that a joint coherence graph is always defined with respect to a dialogue $d$.

The joint coherence graph is initially empty. Each move adds its premises and conclusion as new nodes, after which the edges and coherence values are recalculated according to the definition of $\zeta$ (Definition 5.2.3). In addition, if a

move proposes a norm as an alternative to an earlier proposal for the same goal, we also add the corresponding constraint between the two norms as a new node.

**Definition 8.1.6** *For any dialogue $d$, the* joint coherence graph $\mathcal{J}_d = \langle V_d, E_d, \zeta_d \rangle$ *associated with $d$ is defined as follows:*

- $V_{d_0} = E_{d_0} = \emptyset$ *while* $\zeta_{d_0}$ *is undefined;*
- *For any move* $m = \rho$ *since* $\Gamma$ *:*
  - $V_{(d,m)} = V_d \cup \{\rho\} \cup \Gamma \cup C$, *where:*
    - *if* $m = (O\varphi, r)$ *since* $(B(\varphi \rightarrow \psi), s), (D\psi, t), S$[1] *then* $C = \{(\neg(O\varphi \wedge O\varphi'), f(r, r')) \mid d$ *contains a move with argument* $(O\varphi', r')$ *since* $(B(\varphi' \rightarrow \psi), s'), (D\psi, t'), S$ *such that* $\varphi \neq \varphi'\}$;
    - *otherwise* $C = \emptyset$
  - $\zeta_{(d,m)} = \zeta(\{v, v'\}) \mid v, v' \in V_{(d,m)}$ *from Definition 5.2.3*
  - $E_{(d,m)} = \{(\{v, v'\}) \mid v, v' \in V_{(d,m)}$ *and* $\zeta_{(d,m)}(\{v, v'\})$ *is defined*$\}$

In the Example 7.1.1, the joint coherence graph after the move with the norm proposal

$$m_1 = (O\alpha, 0.9) \text{ since } \{(B(\alpha \rightarrow e), 1), (De, 1), (B(\alpha \rightarrow f), 0.9), (Df, 0.9)\}$$

is as given in Figure 8.1.

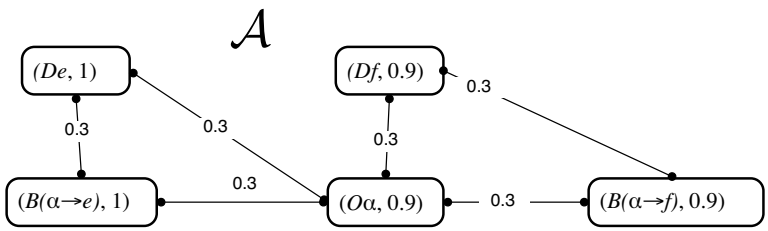


Figure 8.1: Joint coherence graph $\mathcal{J}_d$ with $d = m_1$

Since, the coherence maximisation by agent $b$ classifies the norm $(O\alpha, 0.9)$ from agent $a$into the rejected set as observed in Figure 7.6 in Section 7.4, agent $b$ looks for the highest preferred norm from its accepted set based on the norm generation procedure on Section 7.4 which is $(O\beta, 0.9)$. The support set of the norm is $\{(B(\beta \rightarrow q), 0.9), (Dq, 0.9), (B(\beta \rightarrow e), 1), (De, 1)\}$. Hence, $b$ generates the dialogue move

$$m_2 = (O\beta, 0.9) \text{ Since } \{(B(\beta \rightarrow q), 0.9), (Dq, 0.9), (B(\beta \rightarrow e), 1), (De, 1)\}$$

[1] $S$ is a, possibly empty, set of additional premises.

The elements of the dialogue move $m_2$ are added to the joint coherence graph $\mathcal{J}_{(d,m_2)}$ (Recall that $d, m$ stands for the result of appending $m$ to $d$, i.e., for $d$ as continued by $m$ and hence $\mathcal{J}_{(d,m_2)}$ stands for the joint graph of dialogue $d = m_1, m_2$). Since $\mathcal{J}_{(d,m_2)}$ now has two norms $(O\alpha, 0.9)$ and $(O\beta, 0.9)$ and both realise the same social goal $e$, they cohere negatively in the joint coherence graph. The joint coherence graph $\mathcal{J}_{(d,m_2)}$ is as shown in Figure 8.2.

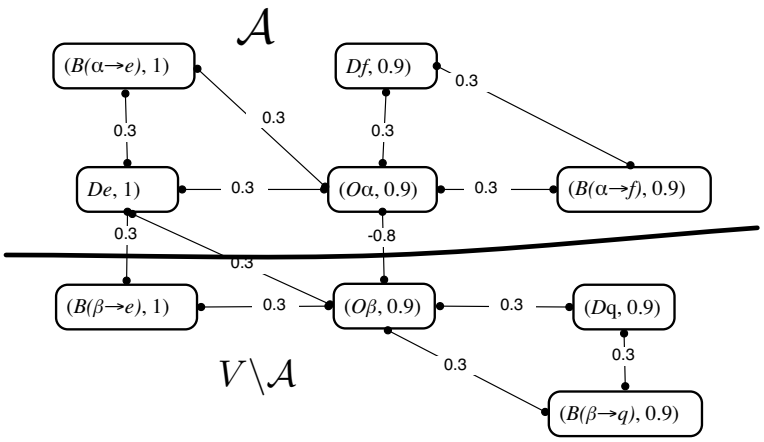


Figure 8.2: Joint coherence graph $\mathcal{J}_d$ with $d = m_1, m_2$ (a coherence maximising partition)

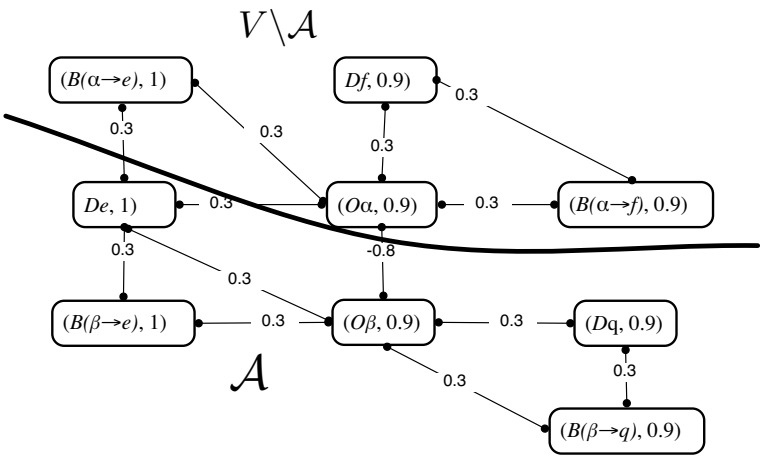


Figure 8.3: Joint coherence graph $\mathcal{J}_d$ with $d = m_1, m_2$ (another coherence maximising partition)

Note that there are at least two coherence maximising partitions that are indistinguishable as shown in Figure 8.2 and in Figure 8.3. They are indistinguishable because, both have equal number of nodes and equal number of links with equal strengths. They also have equal statistics about number of social goals and norm proposals contained in the accepted sets. Hence, any of this can be made the coherence-maximising partition. This indicates that the dialogue so far has resulted in arguments that have equal strength from both agents.

**Computations on Joint Coherence Graph**

Next we discuss certain computations we can perform on the joint coherence graph. These help to review the state of the dialogue and help understand each agent's position in the dialogue so far. By definition of a joint coherence graph, it contains dialogue moves from all participating agents. At any point, if there are more than one coherence maximising partitions, then each agent may prefer one partition over another depending on the elements of its own accepted sets. An agent may prefer a partition depending on whether the norm proposals for social goals it has proposed are part of the accepted set of the partition. This is intuitive as the social goals proposed by an agent are the goals important for the agent. We call such partitions the *preferred partitions* of an agent. As with coherence maximising partitions, there may be more than one preferred partition after any dialogue move. Note also that the criteria to select a unique coherence maximising partition among all the coherence maximising partitions of the joint coherence graph is independent of the preferred partitions of agents.

**Definition 8.1.7** *A partition* $(\mathcal{A}, V \setminus \mathcal{A})$ *of* $\mathcal{J}_d$ *is a* preferred partition $p_a(d)$ *by agent* $a$ *if the accepted set* $\mathcal{A}$ *of* $p_a(d)$ *contains a norm proposed by* $a$ *for each social goal addressed by* $a$ *in* $d$*.* $\mathcal{A}_p$ *denotes the accepted set of a preferred partition* $p_a(d)$*. The set of all preferred partitions of an agent* $a$ *for a dialogue* $d$ *is denoted by* $\mathcal{P}_a(d)$*.*

For the Example 7.1.1, the preferred partition of agent $a$ is the first coherence maximising partition as shown in Figure 8.2 with the accepted set as indicated. The preferred partition $p_b(d)$ of agent $b$ is the second coherence maximising partition as shown in Figure 8.2 with the accepted set as indicated.

So far we have discussed the common shared language and the joint dialogue structure which is the joint coherence graph of the dialogue system. However, these make sense only in the context of an appropriate dialogue protocol. We next define the dialogue protocol for deliberation.

### 8.1.4 Protocol

As discussed earlier, the protocol is for *theory building dialogues*. The protocol enforces relevance and coherence of dialogues in two ways. Initially, a norm proposal must be made for a social goal that triggered the deliberation. Subsequently, each agent must make sure that each dialogue move should improve its position. At any point in the dialogue it is possible to compute the position

of the agents in the dialogue. The position of the agent intuitively corresponds to the degree of success of the agent so far in the dialogue. Hence it is also a measure to evaluate and control the progress of the dialogue. This is an implicit relevance mechanism: argument moves will be chosen such that they improve the speaker's position, which implies that they must somehow relate to what has been said so far. Another important element of the protocol is that as soon as an agent has an improved position, the turn shifts to the other agent, who must then try to have the improved position, and so on. This builds a dialectical element into dialogues that promotes the efficient exploration of all sides of a problem (cf. [Loui, 1998]'s 'immediate-response' disputes). A dialogue ends in agreement when atleast one of both agents' preferred partitions accept the same set of norms.

These dialogue moves and the turn taking rule are based on the evaluations made on the joint coherence graph. The *position* of an agent in a joint coherence graph is determined on the basis of two factors, the ratio of its social goals in the accepted set with respect to all proposed social goals and the ratio of its norm proposals in the accepted set with respect to all of its norm proposals. This is intuitive as the more of an agent's social goals and norm proposals are accepted, then the more it is likely to succeed in the argument. These evaluations are in reference to the preferred partition of an agent. If there are more than one preferred partitions, then the maximum value among evaluations from all preferred partitions is chosen as the position of the agent.

**Definition 8.1.8** *If $G, N_a, G(\mathcal{A}_p), N_a(\mathcal{A}_p)$ respectively define the total number of social goals addressed by any agent, total number of norm proposals by $a$, number of social goals in the accepted set and the number of norm proposals in the accepted set by agent $a$, then the* position $pos_a(d)$ *of an agent in a dialogue $d$ is given by*

$$pos_a(d) = \max_{\mathcal{A}_p, p \in \mathcal{P}_a(d)} \left\{ \frac{\mid G(\mathcal{A}_p) \mid + \mid N_a(\mathcal{A}_p) \mid}{\mid G \mid + \mid N_a \mid} \right\}$$

The notion of *winner of a dialogue $d$* is based on the position of the agents in the dialogue. It is intended to measure the progress made by each agent in the dialogue with respect to the other. If the position of agent $a$ is higher than the position agent $b$ then $a$ has made more progress in the dialogue and is the winner. If this is equal, then there is said to be no winner for the dialogue.

**Definition 8.1.9** *For any two agents $a$ and $b$ in a dialogue $d$, the* winner $\mathcal{W}(d)$ *of the dialogue $d$ is:*

$$\mathcal{W}(d_i) = \begin{cases} undefined & \text{if } pos_a(d) = pos_b(d) \\ a & \text{if } pos_a(d) > pos_b(d) \\ b & otherwise \end{cases}$$

At each stage of the dialogue, an agent must try to be the winner of the dialogue. Initially, when $d = d_0$ the position of both agents is set to 0. Depending

on the position of each agent in the subsequent moves, the *turn taking rule* is defined. If agent $a$ is the current *winner*, then the turn shifts to agent $b$. If there is no winner for $d$, and if $a$ is the agent that made the current move, then the turn shifts to $b$.

In the example in Example 7.1.1, initially, the position of both $a$ and $b$ is set to 0. When agent $a$ makes a move, the position of $a$ becomes $pos_a(d) = \frac{2+1}{2+1} = 1$. The current winner $\mathcal{W}(d_i)$ is agent $a$ as $b$ still has its position as 0. Hence, the turn shifts to agent $b$ which proposes the second norm $(O\beta, 0.9)$. Then the position of $b$ becomes $pos_b(d) = \frac{2+1}{3+1} = 0.75$. And recalculating $a$'s position $pos_a(d) = \frac{2+1}{3+1} = 0.75$. Since, there is no winner at this point, the turn automatically shift to agent $a$ since agent $b$ made the last dialogue move.

$a$ now rejects the norm proposal of $b$ by proposing $(O\neg\beta, 0.8)$ along with the support set for the norm $\{(B(\beta \rightarrow \neg c), 0.8), (Dc, 0.8)\}$ as shown in Figure 8.4. That is,

$$m_3 = (O\neg\beta, 0.8) \; since \; (B(\beta \rightarrow \neg c), 0.8), (Dc, 0.8)$$

. Note that incidentally, this support set along with the norm forms the conflict set for $b$'s norm proposal. Hence, even though there is no direct rejection move in the language, the rejection of a norm or in general an argument can be emulated in this manner.

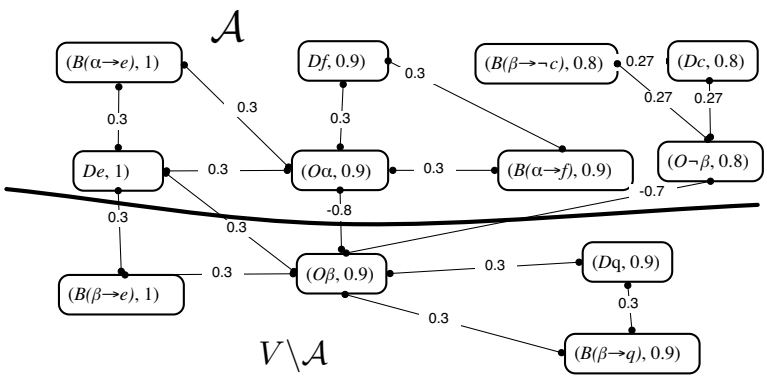


Figure 8.4: $\mathcal{J}_d$ when $d = m_1, m_2, m_3$

Finally, agreement is defined with the following notions. For any partition $p = (\mathcal{A}, V \setminus \mathcal{A})$ of graph $\mathcal{J}$, let $N_a(p)$ denote the set of norms proposed by $a$ belonging to $\mathcal{A}$.

**Definition 8.1.10** *The agents $a$ and $b$* agree *in dialogue $d$ if all focal goals have been addressed in $d$ and there exist preferred partitions $p_a(d)$ and $p_b(d)$ of $\mathcal{J}_d$ such that $N_a(p_a(d)) = N_b(p_b(d))$.*

In other words, the agents agree if they have discussed all focal goals and if they have preferred partitions that contain the same set of norms for all goals

addressed in the dialogue (which may include more goals than just the focal goals, namely, if a move has proposed a new social goal).

In the Example 7.1.1, agent $a$'s position after the dialogue $d = m_1, m_2, m_3$ is $pos_a(d) = \frac{3+2}{4+2} = 0.833$ whereas $b$'s position $pos_b(d) = \frac{1+1}{4+1} = 0.4$. Since $a$ is the winner, it is now $b$'s turn to move. Agent $b$ adds $a$'s proposal (move $m_3$) to its internal coherence graph as in Figure 8.5. It's coherence maximising partition now accepts norms $(O\alpha, 0.9)$ and $(O\neg\beta, 0.8)$ and rejects its own norm proposal $(O\beta, 0.9)$.

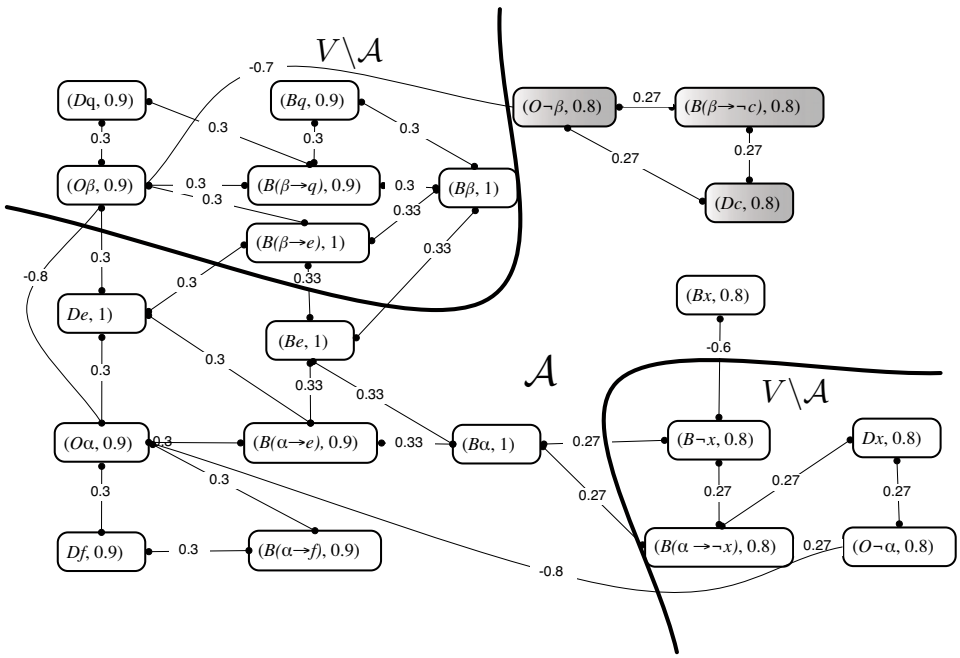


Figure 8.5: Internal coherence graph of agent $b$ when $d = m_1, m_2, m_3$ (agent $a$'s proposal are the shaded nodes)

The highest preferred norm in its accepted set now is $(O\alpha, 0.9)$ which is its candidate for the next norm proposal. Agent $b$ now proposes

$$m_4 = (O\alpha, 0.9) \; \textit{Since} \; \{(B(\alpha \rightarrow e), 1), (De, 1), (B(\alpha \rightarrow f), 0.9), (Df, 0.9)\}$$

which is a repetition of $a$'s first proposal. The joint coherence graph does not get updated with the new proposal as it is a repetition of a previous proposal already incorporated in the graph. However what changes is the preferred partition of $b$ which now is the same as that of agent $a$ as shown in Figure 8.6. The new position of agent $b$ is now $pos_b(d) = \frac{2+1}{4+2} = 0.5$ whereas $a$'s position remains at 0.833.

It is interesting to note that even though both the agents share the same accepted set for their preferred partitions, the positions are not equal. This is because $a$ has made most of the winning proposals and $b$ has not yet made all the norm proposals in the accepted set of the preferred partition. The termination condition which is that all focal goals have been addressed and that $N_a(p_a(d)) = N_b(p_b(d))$. Here $N_a(p_a(d)) = \{(O\alpha, 0.9), (O\neg\beta, 0.8)\}$ whereas $N_b(p_b(d)) = \{(O\alpha, 0.9)\}$. Hence, the termination condition for agreement is not yet satisfied.

Since, it is still $b$'s turn to move, $b$ makes the proposal

$$m_5 = (O\neg\beta, 0.8) \textit{ since } \{(B(\beta \rightarrow \neg c), 0.8), (Dc, 0.8)\}$$

which is the only norm left in the accepted set of $b$'s internal coherence graph. Now the position of $b$ becomes $pos_b(d) = \frac{3+2}{4+2} = 0.833$ equal to that of $a$. Hence, the conditions for agreement according to Definition 8.1.10 is trivially satisfied and the dialogue terminates in agreement.

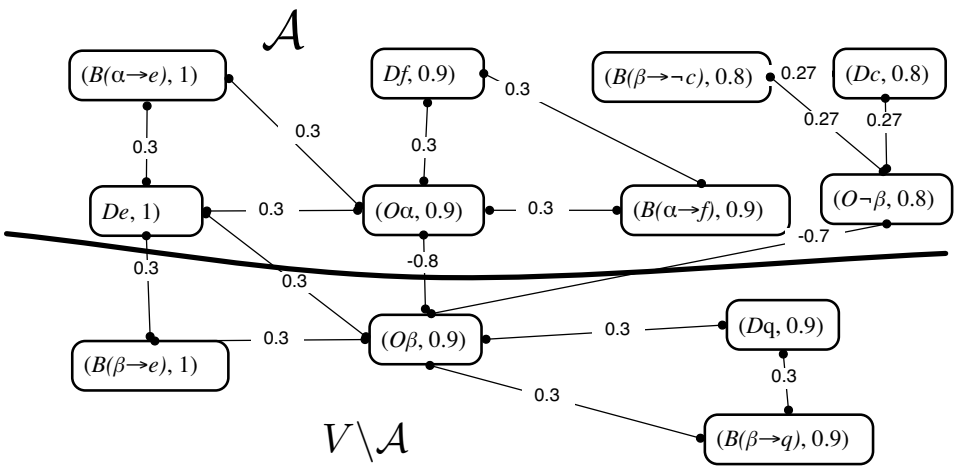


Figure 8.6: Joint graph when $d = m_1, m_2, m_3, m_4$ and the preferred partitions of $a$ an $b$

We now summarise the elements discussed so far to define the protocol. Technically, a *protocol* is a function $Pr$ that assigns to any legal dialogue a set of moves which are its legal continuations. A dialogue is *legal* if any move in it is a legal continuation of the sequence to which it is appended. If $Pr_d = \emptyset$ then $d$ is a *terminated* dialogue.
As explained above, our protocol assumes that each dialogue is against the background of a set $F = \{(D\varphi_1, r_1), \ldots, (D\varphi_n, r_n)\}$ of *focal social goals*. The protocol is then defined as follows.

**Definition 8.1.11** *For any two agents $a$ and $b$ and for any dialogue $d$, $m = (a, x) \in Pr_d$ iff:*

1. *a is the agent to move in d;*
2. *if $d = d_0$ then $x$ is of the form $(O\phi, r)$ since $(B(\phi \rightarrow \psi), s), (D\psi, t), S$ where $(D\psi, t)$ is a focal goal and $S$ is the set of the remaining elements in the support set of $(O\phi, r)$;*
3. *$E_{(d,m)}$ contains positive support links from each premise of $x$ to its conclusion;*
4. *if $\mathcal{W}(d_i)$ is undefined then*
   (a) *either $m$ is the first move in the dialogue or,*
   (b) *$b$ made the last move in $d$.*
5. *$d$ contains no move $(a, x)$;*
6. *the agents do not agree in $d$.*

The interpretation for each of the rules are given below.

1. Rule (1) is obvious. An agent who has a norm to propose makes the first dialogue move. If both agents have norms to propose, then this selection is arbitrary.
2. Rule (2) says that each discussion starts with a proposal for a norm that (if complied with) achieves some focal social goal. Each subsequent move may be an argument of any form, as long as it respects the remaining protocol rules.
3. Rule (3) says that each move must be an argument in that in the resulting joint coherence graph, the premises of the move must positively cohere with its conclusion.
4. Rule (4) says that a move can only be made if the speaker is not the winner of the dialogue so far, or it is the first move in the dialogue or the speaker is not the agent to make the last move. This is the turn taking rule explained previously.
5. Rule (5) prevents an agent from repeating his own moves.
6. Rule (6) makes sure that a dialogue terminates after the agents have reached agreement.

## 8.2 Comparison with Other Argumentation Systems

We now make a detailed comparison of our protocol with logic-based protocols for reaching agreement over action. The most detailed proposal we know of is that of Atkinson [Atkinson, 2005b], who derives a dialogue protocol from

an extended version of Walton's [Walton, 1996] argument scheme for justifying actions and its critical questions. Atkinson's critical questions (CQ) generate several ways to attack an argument that instantiates the practical reasoning scheme (e.g. alternative ways to reach the same goal, negative effects on other goals, the beliefs on which the argument is based are false, and so on). Here we evaluate to what extend similar attacks of the various types can be launched against norm proposals generated with the first bridge rule. Note first that we have a restricted domain ontology in that unlike Atkinson we do not distinguish between goals and values, between truth and possibility and between circumstances and actions. Let it be of the form $O\phi$[2] *since* $(B(\phi \rightarrow \psi)), D\psi$. All these simplifications are meant to focus on the essence of our proposoal, which is its use of the coherence mechanism. These simplifications make that only a number of Atkinson's critical questions are relevant for our model (since we do not distinguish between values and goals, we have replaced Atkinson's term 'value' in CQs 9 and 10 by 'goal'):

Let us see to what extent our protocol allows dialogue moves to be moved as arguments in reply to an application of the first bridge rule (Rule 1 of Section 7.4).

- *CQ1: Are the believed circumstances true?* Since we model the deliberation of norm givers on how to regulate a domain instead of the reasoning of agents on what to do in concrete situations, this question is irrelevant for us.

- *CQ2: Assuming the circumstances, does the action have the stated consequences?* This can be addressed with an argument for conclusion $(B(\neg(\phi \rightarrow \psi)))$. This move will introduce a negative coherence link between this conclusion and the original belief $(B(\phi \rightarrow \psi))$.

- *CQ5: Are there alternative ways of realising the same consequences?* This can be formulated with an alternative application of Rule 1: $O\phi'$ *since* $(B(\phi' \rightarrow \psi)), D\psi$. Combined with the constraint $\neg(O\phi \wedge O\phi')$ introduced by this move adding a negative support link between $O\phi$ and $O\phi'$.

- *CQ9: Does doing the action have a side effect which demotes some other goal?* We can express this by an application of the Rule 2. This adds a node $O\neg\phi$ to the joint coherence graph, which negatively coheres with the node $O\phi$.

- *CQ10: Does doing the action promote some other goal?* We can express this by applying the first bridge rule Rule 1 of Section 7.4 to the other goal, resulting in another argument for the same norm. As shown above, this normally improves the speaker's position and thus naturally models accrual of arguments.

- *CQ11: Does doing the action preclude some other action which would promote some other goal?* This corresponds to the situation that we have

[2]Here the grades are ignored for convenience.

$(B(\phi \rightarrow \neg\psi))$ and $(B(\psi \rightarrow \chi))$ and $D\chi$. Roughly, we can only express this if $\psi \rightarrow \chi$ is necessarily true, i.e., true in all possible worlds: then the argument for $O\phi$ can be countered with an argument for $O\psi$ applying the first bridge rule and further extended to $O\neg\phi$: then $O\phi$ and $O\neg\phi$ negatively cohere in the joint coherence graph.

Concluding, given our restricted domain ontology, our model essentially allows for all argument moves and critical questions proposed by Atkinson; a possible advantage of our approach over Atkinson's is a natural way to model accrual of alternative arguments for the same norm (which is arguably more natural than [Bench-Capon and Prakken, 2006]'s logic-based model of accrual).

## 8.3 Discussion

In this chapter, we have discussed the interaction between autonomous normative agents and the normative system. In particular, we discussed the dynamic aspects of this interaction and discussed a mechanism by which a groups of agents can self regulate their normative system, propose and establish operational norms based on a set of social goals. A dialogue system was proposed which uses coherence-driven argumentation to guide the deliberation process. Unlike traditional argumentation systems, the joint coherence graph which is the common dialogue structure that agents build during the deliberation both evaluates and controls the progress of the dialogue. Further, as this dialogue system is coherence-driven, modelling uncertainty becomes natural. More importantly, as stated in the introduction, a coherence-driven argumentation is meant to provide flexibility in stating arguments, introducing support and attack relations as a matter of degree, and accepting those maximally coherent arguments that might have inconsistencies.

In the next chapter, we try to understand the kind of agents that can be modelled with coherence maximisation. As rational agents are one of the interesting kind, we compare and contrast coherence maximisation against utility maximisation and show that utility maximising agents are essentially coherence maximisers of a particular type of coherence graphs. However, we go one step further to show that, not only rational agents of a strict economic sense, but agents that operate based on other values can also be represented by coherence maximisation. This also highlights the dynamic nature of coherence maximisation, where different graphs can be joined dynamically which reflect the situational changes or changes in beliefs thus resulting in a dynamic computation of preference ordering.

# Chapter 9

# Coherence: When is it Right?

So far we have been interested in the process of coherence maximisation, and designing agents that are coherence-driven. We also looked at coherence maximisation of a group of agents as a way of reaching consensus. In this chapter, we take a step back and analyse coherence maximisation in a broader context. Our aim is to understand whether it is a rationally interesting behaviour (in the neo-classical economic sense) to be coherence-driven and what kind of agents can be modelled as coherence-driven. To answer, we show that the traditional notion of utility maximisation—which is an accepted rational behaviour— can be emulated using coherence maximisation. In addition, we show that other types of agents which are not strict utility maximisers in the economic sense can also be modelled using this approach.

We propose what we call *utility coherence graphs* along with coherence maximisation to emulate the behaviour of utility functions. We first discuss the particular interpretation of rationality we use in this chapter, and discuss the assumptions on the preference ordering in Section 9.1. In Section 9.2, we prove that *the maximum element of a preference ordering is the same as that found by coherence maximisation over a corresponding utility coherence graph.* We illustrate this by modeling the *ultimatum game* with a utility coherence graph. Further, in Section 9.3, we argue that utility coherence graphs have marked advantages over utility functions, in that they blend well with the internal representation and reasoning of cognitive agents. To model an altruistic agent using a traditional utility maximising function, one needs to change the preference function. However, it is not clear, how this preference function can be computed. We show that leaving the preference ordering as such, we can incorporate other beliefs of the agents to model altruism, reputation, or the like. This means coherence maximisation using coherence graphs can adapt to changing preferences and model different types of agents. We illustrate this by modelling an agent that wishes to follow a social custom of being fair, keeping the original

preference ordering as such. This also shows the generality of the coherence-based approach, which is better suited to model different senses of rationality. We conclude with a discussion in Section 9.4.

## 9.1 Background

In this section, we make clear the interpretation of rationality used for the purpose of this chapter. We also discuss preference relations which are basic in defining utility functions. The idea is to take the same assumptions of a utility function in defining the utility coherence graphs.

### 9.1.1 Rationality

In neo-classical economics, rationality is idealised to decisions that are optimal by maximising utility or profit, for realising goals of an adaptive system [Simon, 1969]. In this chapter, we confine to this utility-maximising idealisation of rationality. A basic assumption while modelling a rational agent is the existence of an a priori ordering of preferences based on a meaningful measure of utility. Utility functions are practical representation mechanisms of preferences because we can apply standard optimisation techniques.

However, we encounter certain difficulties when using a utility function to model the behaviour of a rational agent. Firstly, an autonomous agent chooses to pursue an action by considering various factors influencing its decision, such as the norms of a society it is part of, its reputation, the context of the action, altruism, etc. Those supporting a utility-based approach often claim that preferences should include all such considerations. However, it is very hard to compute such a preference, and we do not share the view that they are basic. On the contrary, preferences are the consequence of deliberation, reasoning, and other complex cognitive processes built upon basic cognitions.

The second and the most important difficulty is to measure the influence of new information on the preference ordering. Utility functions are either static or do not have transparent computational mechanisms to readjust the preferences. Both these difficulties arise mostly because utility functions do not blend well with the representation of the cognitions of an agent such as its goals (desires), actions (intentions), plans, or even beliefs. In our view, preferences are indeed a consequence of the interaction of all the cognitions, they are dynamic, possibly uncertain and imprecise. Therefore, a simplistic static linear ordering over outcomes, as a utility function establishes, falls short for complex reasoning agents. Coherence maximisation offers a more global perspective to decision making, considering how a cognition (belief, goal, action, or commitment) gets supported and gives support to other cognitions. In the context of coherence, preferences over outcomes are the consequence of the interactions among cognitions.

### 9.1.2 Preference Relation

Preferences are relevant when it is necessary to examine the behaviour of an individual who must choose from a set of outcomes $O$. A preference relation basically describes an order of preference of an agent among a set of alternative outcomes.

**Definition 9.1.1** *A* preference relation $\succcurlyeq$ *on a finite set of outcomes* $O$ *is a total pre-order on* $O$*, i.e., for all* $o, p, q \in O$*,*

- $o \succcurlyeq o$ *(*$\succcurlyeq$ *is reflexive)*
- *if* $o \succcurlyeq p$ *and* $p \succcurlyeq q$ *then* $o \succcurlyeq q$ *(*$\succcurlyeq$ *is transitive)*
- $o \succcurlyeq p$ *or* $p \succcurlyeq o$ *(*$\succcurlyeq$ *is complete)*

When $o \succcurlyeq p$ we say that $o$ *is at least as preferable as* $p$. We write $o \sim p$ when $o \succcurlyeq p$ and $p \succcurlyeq o$, and $o \succ p$ when $o \succcurlyeq p$ but $o \not\sim p$. The down-set of $o \in O$ is $\downarrow o = \{p \in O | o \succcurlyeq p\}$. Further, due to the property of $\sigma$ given in Equation (3.1) (see Section 3.1), we shall require the set $O$ to have at least three outcomes. Here we consider only those preference relations on $O$, that have a maximum.

## 9.2 Utility Coherence Graphs

As mentioned in the introduction, a utility function is used to assign a numerical value to the preference ordering, so that we can employ numerical techniques to maximise utilities in order to select an outcome that is most preferred. As stated in the introduction, a coherence graph is a better choice for a mathematical formulation of preferences as it blends into the representation of the agents, and coherence maximisation helps agents retain their autonomy. The aim of this section however is to prove that, given a preference relation, there exists a utility coherence graph which effects the behaviour of a utility-maximising function. Further, we also show how such a graph would look like. We first discuss the conditions for a utility coherence graph in the form of a lemma and then use this lemma to prove the theorem which relates utility maximisation with coherence maximisation.

We now state a lemma that defines the necessary conditions under which a graph $g$ will be a utility coherence graph. The first condition states that $o \succ p$ with respect to a preference relation, if and only if, when accepted in $g$, their respective total strengths preserve the ordering. In other words, it is more coherent to accept a more preferred outcome to a less preferred one. The second condition states that, for the maximum in $O$, it is more coherent to accept the maximum alone than accepting it with other less preferred outcomes.

To prove the lemma, we define a coherence graph as follows. We take the set of outcomes as nodes of the graph. Since the outcomes are mutually exclusive, it is natural to assign negative coherences between them. The degrees of incoherence depend on the number of equivalence relations between the outcomes, and

to satisfy the second condition of the lemma below, the degrees should decrease exponentially as less preferred nodes are linked in the graph. Hence, we use the cardinalities of down-sets to define the degree of coherence between two nodes $a$ and $b$ as $-|O|^{|\downarrow a|+|\downarrow b|}$. However, there may exist other ways to determine the degrees of incoherence that satisfy the conditions of the lemma.

**Lemma 9.2.1** *Let $O$ be a finite set of outcomes such that $|O| \geq 3$, and let $\succcurlyeq$ be a preference relation on $O$. Then there exists a coherence graph $g$ such that,*

*a) for all $o, p \in O$, $o \succ p$ if and only if $\sigma(g, \{o\}) > \sigma(g, \{p\})$;*

*b) if $o$ is the maximum of $O$, for all $P \subseteq O$ such that $o \in P$ and $P \neq \{o\}$; then $\sigma(g, \{o\}) > \sigma(g, P)$.*

**Proof 9.2.2** *Let $g = \langle V, E, \zeta \rangle$ be a coherence graph such that*

- $V = O$
- $E = \{\{o, p\} \mid o, p \in O\}$
- *for all $o, p \in O$ $\quad \zeta(\{o, p\}) = -|O|^{|\downarrow o|+|\downarrow p|}$ (Note that, for the purpose of not cluttering the equations in the proof, we henceforth denote $|\downarrow a|$ in the exponent simply as $a$ ).*

*That is, in $g$ all nodes are incoherent between each other. In such a graph, given a partition $(\mathcal{A}, V \setminus \mathcal{A})$ of $V$, the set of satisfied constraints is, by Definition 3.1.2, $C_{\mathcal{A}} = \bigcup_{v \in \mathcal{A}} \bigcup_{w \in V \setminus \mathcal{A}} \{\{v, w\}\}$. Consequently, by Definition 3.1.3,*

$$\sigma(g, \mathcal{A}) = \frac{\sum_{v \in \mathcal{A}} \sum_{w \in V \setminus \mathcal{A}} |\zeta(\{v, w\})|}{|E|} \ .$$

*Since in $g$, $V = O$ and, for all $o, p \in O$, $|\zeta(\{o, p\})| = |O|^{o+p} = |O|^o \cdot |O|^p$, we have that*

$$\sigma(g, \mathcal{A}) = \frac{\sum_{o \in \mathcal{A}} \sum_{p \in O \setminus \mathcal{A}} |O|^o \cdot |O|^p}{|E|} \ .$$

*a) For all $o, p \in O$, $\sigma(g, \{o\}) > \sigma(g, \{p\})$ if and only if*

$$\sum_{q \in O \setminus \{o\}} |O|^o \cdot |O|^q > \sum_{q \in O \setminus \{p\}} |O|^p \cdot |O|^q \ .$$

*Splitting the summations to extract the common term $|O|^o \cdot |O|^p$ and cancelling it out, the inequality is equivalent to*

$$|O|^{o+p} + |O|^o \cdot \sum_{q \in O \setminus \{o,p\}} |O|^q \quad > \quad |O|^{p+o} + |O|^p \cdot \sum_{q \in O \setminus \{o,p\}} |O|^q \ ,$$

*which is equivalent to*

$$|O|^o \cdot \sum_{q \in O \setminus \{o,p\}} |O|^q \quad > \quad |O|^p \cdot \sum_{q \in O \setminus \{o,p\}} |O|^q \ .$$

*Since* $|O| \geq 3$, *we have that* $\sum_{q \in O \setminus \{o,p\}} |O|^q > 0$, *and consequently, the above inequality is equivalent to* $|O|^o > |O|^p$, *which holds if and only if* $|\downarrow o| > |\downarrow p|$, *and if and only if* $o \succ p$.
*b) Let* $o$ *be the maximum in* $O$, *and let* $P \subseteq O$ *such that* $o \in P$ *and* $P \neq \{o\}$. *By Definition 3.1.3,*

$$\sigma(g, \{o\}) > \sigma(g, P) \text{ if and only if } \sum_{q \in O \setminus \{o\}} |O|^o \cdot |O|^q \quad > \quad \sum_{p \in P} |O|^p \cdot \sum_{q \in O \setminus P} |O|^q \ .$$

*Taking the common terms from both the sides, we get,*

$$|O|^o \cdot \Big( \sum_{p \in P \setminus \{o\}} |O|^p + \sum_{q \in O \setminus P} |O|^q \Big) \quad > \\ |O|^o \cdot \sum_{q \in O \setminus P} |O|^q + \sum_{p \in P \setminus \{o\}} |O|^p \cdot \sum_{q \in O \setminus P} |O|^q \ .$$

*Cancelling the common term* $|O|^o \cdot \sum_{q \in O \setminus P} |O|^q$ *on both the sides, we have*

$$|O|^o \cdot \sum_{p \in P \setminus \{o\}} |O|^p \quad > \quad \sum_{p \in P \setminus \{o\}} |O|^p \cdot \sum_{q \in O \setminus P} |O|^q \ .$$

*Since* $\sum_{p \in P \setminus \{o\}} |O|^p > 0$, *the above inequality is equivalent to*

$$|O|^o \quad > \quad \sum_{q \in O \setminus P} |O|^q \ .$$

*To prove this inequality, let* $p \in P$ *and* $p \neq o$. *Then we have,*

$$\sum_{q \in O \setminus P} |O|^q \quad \leq \quad \sum_{q \in O \setminus \{o,p\}} |O|^q \quad .$$

*Since* $o$ *is the maximum, for all* $q \in O$ *and* $q \neq o$, $|O|^q \leq |O|^{o-1}$. *Consequently,*

$$\sum_{q \in O \setminus \{o,p\}} |O|^q \leq (|O| - 2) \cdot |O|^{o-1} < |O| \cdot |O|^{o-1} = |O|^o \ .$$

□

An example of a utility coherence graph is in Figure 9.1. Note that the coherence maximising partition has the most preferred outcome as the accepted set, $\mathcal{A} = \{o_1\}$.

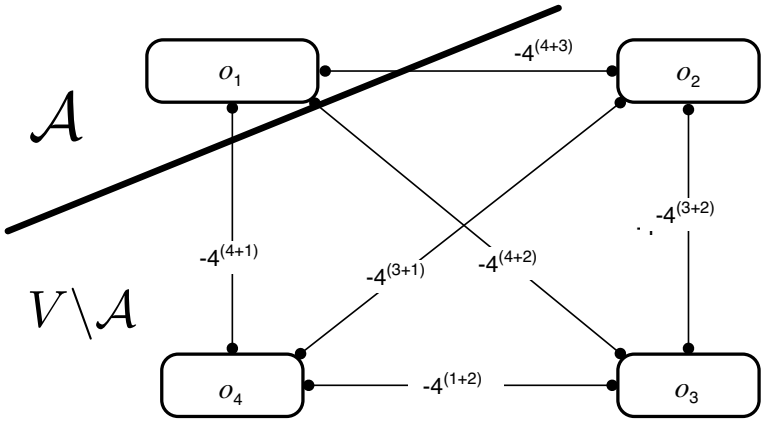


Figure 9.1: An example of a utility coherence graph given $o_1 \succ o_2 \succ o_3 \succ 0_4$

We now define the theorem that may be summarised by saying that, there exists a coherence graph with certain properties that respects an a priori preference relation. The theorem makes it possible to formulate preference relations as coherence graphs which blends into the representation of a rational agent and can model autonomous agent behaviour.

**Theorem 9.2.3** *Given a finite set of outcomes $O$ such that $|O| \geq 3$, and a preference relation $\succcurlyeq$ on $O$, there exists a coherence graph $g$ such that, $o$ is the maximum in $O$ with respect to $\succcurlyeq$ if and only if $\{o\}$ is the accepted set of a coherence maximising partition of $g$.*

**Proof 9.2.4** *We proceed by contradiction. First let us assume that $o$ is not the maximum element of $O$ and that $\{o\} = \arg\max_{Q \subseteq O} \sigma(g, Q)$. However, if $o$ is not the maximum in $O$, then there exists a $p \in O$ such that $p \succ o$. Then, by Lemma 1 a), we have $\sigma(g, \{p\}) > \sigma(g, \{o\})$, which is a contradiction.*

*Now let us assume that $o$ is the maximum in $O$ and $\{o\} \neq \arg\max_{Q \subseteq O} \sigma(g, Q)$. Then, there exists a $P \subseteq O$ such that $P = \arg\max_{Q \subseteq O} \sigma(g, Q)$ and both $P \neq \{o\}$ and $O \setminus P \neq \{o\}$, because of the property of $\sigma$ given in Equation (3.1) stated in Section 3.1.*

*On the one hand, if $o \in P$, then, by Lemma 1 b), we have $\sigma(g, \{o\}) > \sigma(g, P)$, which is a contradiction. On the other hand, if $o \notin P$, then $o \in O \setminus P$. By Lemma 1 b), $\sigma(g, \{o\}) > \sigma(g, O \setminus P)$. By Equation (3.1) , $\sigma(g, \{o\}) > \sigma(g, P)$, which is a contradiction.* □

We take the example of the *ultimatum game* to demonstrate the theorem just discussed. The ultimatum game is a game often played in economic experiments in which two players interact to decide how to divide a sum of money that is

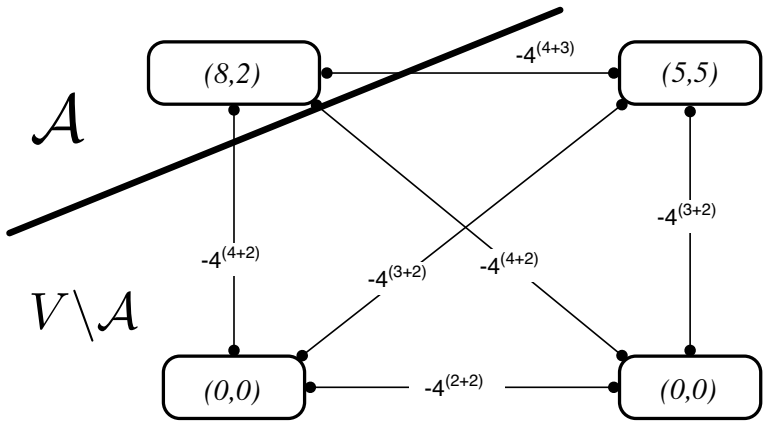


Figure 9.2: The utility coherence graph of *proposer* in ultimatum game

given to them. The first player (*proposer*) proposes how to divide the sum between the two players, and the second player (*responder*) can either accept or reject this proposal. If the responder rejects, neither player receives anything. If the responder accepts, the money is split according to the proposal. The game is played only once so that reciprocation is not an issue. Though every numerical value is possible for the offer, to simplify the representation, we here assume two distinct offers 2 representing unfair offers and 5 for fair offers given the money received is 10. The outcomes of the game and their utilities for each player is represented in Table 9.1.

| | *Accept* | *Reject* |
|---|---|---|
| $Fair$ | $(5,5)$ | $(0,0)$ |
| $Unfair$ | $(8,2)$ | $(0,0)$ |

Table 9.1: Ultimatum game outcomes modelled in game theory

To model the proposer using utility coherence graphs, the proposer has the preference ordering on the outcome $(8,2) \succ (5,5) \succ (0,0) \sim (0,0)$. The utility coherence graph for the proposer corresponding to its preference ordering is given in Figure 9.2. A indicated in the figure, we can see that the highest preferred outcome coincides with the accepted set of the coherence maximising partition in the corresponding utility coherence graph.

## 9.3 Dynamism in Preference Ordering

By proving the theorem stated in this chapter, we have shown that coherence maximisation over a utility coherence graph can emulate the behaviour of a

utility maximising function. We see this as the first step in relating the concept of coherence and that of rationality as understood traditionally. However, the utility maximising interpretation of rationality has been questioned for reasons stated in the beginning of this chapter. For example, the experimental results of the ultimatum game suggests that the behaviour of human subjects often tend to deviate from this utility maximising strategy. That is, the proposer more often choses a fair offer (represented in this example as $(5, 5)$, however it can be anything around these values) than the utility maximising offer of $(8, 2)$. Researchers who have studied these experiments have tried to explain this phenomena with the help of concepts like *fairness*, *generosity*, *altruism*, *reputation*, or even *social custom*. Whatever may be these reasons, it is easy to see that accounting for such reasons has no formal representation in the context of a traditional utility maximising function. That is, preference ordering is often defined like a black box which takes into account all the influences that contribute to generating the preferences.

Even though it is hard to imagine such a preference ordering function, we can produce such an effect by joining a utility coherence graph (of an incomplete preference ordering) with the rest of the coherence graphs of an agents cognition. For the example of the ultimatum game, to model a proposer who desires to follow the social custom of being fair, we can merge the utility coherence graph with the assumed preference ordering and part of its cognitive coherence graph as discussed in Section 5.2.1. To keep the graph simple, the only nodes that are added are the belief to be fair and its consequences. Note that we now use belief coherence graph of the agent (refer to Section 5.2.1). We make a further change in the nodes of the utility coherence graphs so that they now represent beliefs, and not simply numerical values. Hence $(5, 5)$ will now be represented as $(B(5, 5), 1)$ which states that the belief that the outcome is $(5, 5)$ is 1. For simplicity again, we assume the degrees on the outcome as all equal to 1. Note also that, to eliminate biases, we normalise the edge weights of the utility coherence graph to fall between $-1$ and 1 before introducing the new information, dividing by the largest possible absolute value of an edge in a utility coherence graph, which is $|O|^{|O|+|O|}$ where $O$ is the set of nodes (in this example it is $4^8$). We use $(B5, 1)$ to denote the belief that *the offer is* 5 is 1. Since in this example, we represent fairness with the value 5, the two possible outcomes are either the responder accepting the fair proposal leaving both with $(5, 5)$ or rejecting, leaving both to earn $(0, 0)$. That is, $(B5, 1) \rightarrow (B((5, 5) \vee (0, 0)), 1)$. Since these outcomes are mutually exclusive, we also have $(B((5, 5) \wedge (0, 0)), 1) \rightarrow \bot$. A closure of the theory is given in Table 9.2.

Consequently, we have the following deductions:

$$(B5, 1) \rightarrow (B(5, 5) \vee (0, 0), 1), (B(5, 5), 1), (B5, 1), (B((5, 5) \wedge (0, 0)), 1) \rightarrow \bot \quad \vdash \quad (B\neg(0, 0), 1)$$

$$(B5, 1) \rightarrow (B(5, 5) \vee (0, 0), 1), (B(0, 0), 1), (B5, 1), (B((5, 5) \wedge (0, 0)), 1) \rightarrow \bot \quad \vdash \quad (B\neg(5, 5), 1)$$

| *Theory* | $\mathcal{A}$ |
|---|---|
| $\mathcal{T}^{\bullet}{}_{B}$ | $(B5,1) \rightarrow (B((5,5) \vee (0,0)),1)$, |
| | $(B((5,5) \wedge (0,0)),1) \rightarrow \bot$ |
| | $(B5,1),(B(5,5),1),(B(0,0),1)$ |
| | $(B(8,2),1),(B(0,0),1),(B((5,5) \vee (0,0)),1)$ |
| | $(B((5,5) \wedge (0,0)),1)$ |

Table 9.2: Belief theory under sub-formula closure of the *proposer*

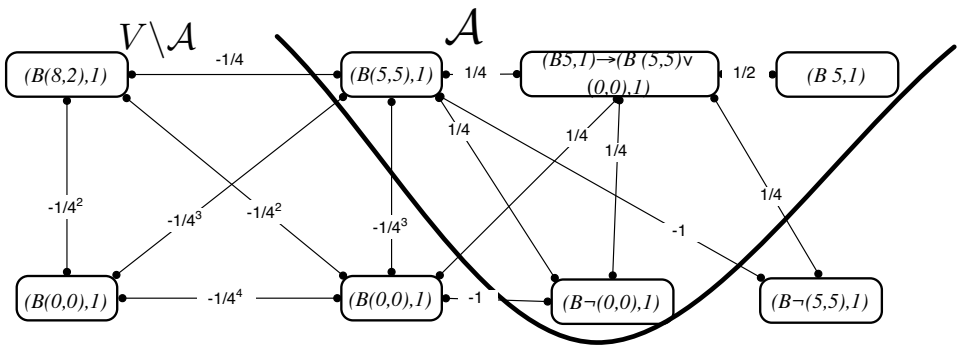


Figure 9.3: The joint coherence graph of *proposer* with $(B5,1)$ added.

$$(B(0,0),1),(B\neg(0,0),1) \quad \vdash$$
$$(B(5,5),1),(B\neg(5,5),1) \quad \vdash$$

As in Figure 9.3, the coherence maximising partition has been changed with the accepted set now containing the outcome $(B(5,5),1)$.

## 9.4 Discussion

In this chapter, we have analysed coherence maximisation in the context of the economic interpretation of rationality. Our analysis not only prove that coherence maximisation can model this traditional sense of rationality, but can go beyond it to model different types of agents other than the economic utility maximisers. We further see that, one of the main limitations of a utility maximsing function which is the dependence on an atomic preference ordering can be overcome with coherence maximisation. By merging different cognitive coherence graphs of an agent with that of a utility coherence graph, what we are doing in effect is finding a new preference ordering that is in sync with the agents cognitions.

With this generalised view of coherence maximisation, we conclude the main contributions of this book. In the next chapter on conclusions, we summarise the main findings, and provide certain insights into the future work, especially on the kind of problems that can be solved with this framework. We end the book by analysing the theory of coherence in the context of some of the well known philosophical theories.

# Part IV

# Conclusion and Future Directions

# Chapter 10

# Concluding Remarks

*"One never notices what has been done;*
*one can only see what remains to be done."*
Marie Curie (1867 - 1934)

In retrospect, working on this book has been an exploration into the unknown. The overwhelming feeling though is not that of a finished journey, but one that is ready to begin. This book in summary has done just that, ensuring that the journey is worthwhile. The guiding theme for this exploration has been the theory of coherence and finding out how useful it is as a motivational driver for autonomous agent reasoning. A more grandeur motivation for the book was to attempt at designing a MAS that can self-regulate. In placing our final remarks, let us use the notion of coherence and its effect on agent behaviour evolution and on evolution of self regulatory systems. In the journey forward, there are many interesting open directions waiting to be explored. Let us make short remarks on some of the most prominent research paths that lay ahead and are worthy of attention. We conclude with one of the kind of applications where we envision coherence-driven agents to be particularly suitable.

## 10.1 Autonomous Agent Reasoning

As stated in the introduction, a primary motivation for this book was to expand the current scope of agents that are currently closer to simple pieces of software with a pre-designed range of behaviours and that lack much of the flexibility conceived in the concept of autonomous agents. In particular, some of the characteristics we intended to introduce in agents in order to increase flexibility are:

1. ability to generate motivations based on one's own cognitive elements
2. ability to reason autonomously and resolve conflicts considering relevant cognitive elements

3. ability to capture uncertainty in the world model

4. ability to adapt goals and actions to situational changes and changes in cognitive elements

In Chapter 5, we have proposed a coherence-based agent architecture that generates agent motivations and actions based on a notion of coherence derived from the theory of coherence, a cognitive theory. As discussed throughout this book, this enables coherence-driven agents to take decisions that best fit their cognitive elements considered as a whole. Hence, motivations are generated or selected from a set of cognitive elements of agents and, as coherence maximisation is computed based on the entire set of cognitions, it is made sure that a coherence-driven agent reasons autonomously and resolves conflicts based on a global maximisation. In this book, coherence-maximisation always took into account the entire set of cognitive elements (subformula-closed theory presentation) of an agent. However, not all the cognitive elements may be *relevant* for resolving a conflict or deciding upon an action. Sometimes, using the entire set of cognitive elements can undermine the influence of relevant cognitive elements. We have not looked in sufficient detail this problem of determining a relevant subset of the theory presentation or the influence of *context* in decision making based on coherence-maximisation. To some extend, the problem of context is solved by the very method of computing coherence values between pairs of cognitive elements. That is, if two theory elements are not related deductively, then there is no edge connecting them. Hence, a natural context is defined as the set of those edge connected elements defined by a coherence function. However, more work is needed to propose a formal notion of context in coherence graphs.

On the formal side, we have made sure that the architecture preserves the properties of verifiability and reliability that is present in BDI family of agent architectures. Due to the formal techniques employed in generating coherence graphs and computing coherence functions, and due to preserving as much as possible of the BDI architecture, we have ensured that we can verify these properties (Chapter 3 and Chapter 4). Finally, by representing graded cognitions, a coherence-driven agent is able to better capture uncertainty in the world model (Chapter 5).

As an anecdote to the main themes in this book, we have strayed a bit to open fields in search for a deeper understanding of rational agents. What we did find in the artificial intelligence and MAS literature is a definition which equates rational agents to utility maximisers. However, as we see it, utility maximisers are simply one of the kinds of agents and do not represent the entire spectrum of cognitive agents. Coherence maximisation if chosen as a motivational drive has the capability to adapt to different personality traits depending on the cognitive elements present in the agent theory. One of the most important differences between the two approaches is that while utility maximisation assumes the existence of an a priori preference ordering, coherence maximisation is able to compute a realistic preference ordering considering the constraints that exist among the cognitive elements of an agent. In addition, coherence maximisers are

utility maximisers and take decisions based on maximising utility (coherence), thus satisfying the necessary conditions of rational behaviour.

In Chapter 9, however, we pose a different question. That is, whether coherence maximisation can emulate rationality with a traditional definition of rational agents as utility maximisers. The theorem which states that *the maximum element of a preference relation is exactly the same as that found by coherence maximisation over the corresponding utility coherence graph* (Theorem 9.2.3) affirms that coherence maximisation can be used to simulate rational agent behaviour. Later in the chapter, we see that coherence-based approach has marked advantages over a utility maximisation in that the former is able to maintain a dynamic preference ordering which reflects the changes in the knowledge base of an agent. In addition, a utility coherence graph can be merged with other coherence graphs of an agent and achieve behaviour traits such as altruism, norm conformity etc. This is so because a coherence maximisation over such a composition of graphs may select actions which exhibit altruistic or norm abiding agent behaviour.

Concluding from all of the above remarks, we can now claim with sufficient grounds that coherence maximisation is indeed a powerful motivational mechanism to model rational agents in its many flavours.

## 10.2 Normative MAS

A second motivation for this book is to extend the envisioned flexibility in agent architecture to agents situated in regulated environments. This motivation is derived from the kind of developments that MAS witnessed during the last decade, to arrange agent communities under organisations or regulated environments. Among the many benefits, such regulated environments help contain the complexity with the introduction of normative artefacts. Even though the regulated systems are conceived by acknowledging the autonomy of participating agents, in many instantiations such as in electronic institutions [Sierra et al., 2004], regulations are strictly enforced. That is, an agent's autonomous behaviour cannot be enacted in an electronic institution. As stated in the introduction, the presence of autonomous behaviour both helps agents and institutions to adapt over time. In this respect, some of the characteristics we intended to further introduce in agents in order to increase flexibility in a regulated context are:

1. ability to reason autonomously and resolve conflicts considering cognitive elements due to personal motivations and due to norms of regulatory institutions an agent is part of.

2. ability to self-regulate by communicating, deliberating and adapting norms to situational changes and common goals.

Introducing characteristics of a regulatory environment in an agent can be achieved by incorporating normative reasoning in the agent. In Chapter 7, we have discussed this by extending the coherence-based agent architecture to include norms. This was straightforward since coherence-driven agents are based

on the multi-context BDI architecture, and extending the architecture with norms could be done simply by adding a context corresponding to norms and connecting the norm context to other contexts through the use of bridge rules.

Normative reasoning not only includes reasoning about norm compliance in a specific situation, but other capabilities such as the ability to make norm proposals and reason about norm proposals of others motivated by social or private goals. A basic assumption we use here is that an unsatisfied social goal is one of the reasons for proposing a norm that would enable realising the social goal. This is similar to the practical syllogism used for intention generation. We have used coherence as a criteria both to select among possible norms and to reason about norm proposals of others.

Thus coherence-driven agents in regulated environments would consider norms at the level of other cognitive elements and integrate norms in a seamless manner. This kind of integration makes sure that agents understand the effects of norms on their goals and take a rational decision based on maximisation of coherence. This is different from setting static priority rules or other static typing of agents into norm abiding or self interested categories. In this case agents lack an understanding of the effects of norms on their cognitive elements. This also differs from reasoning based on sanctions even though the effects of sanctions can be considered in this architecture.

Normative agents further need to collaborate and deliberate on normative issues to be able to self-regulate a MAS constituted by them. We have considered only one aspect of deliberation in this book, specifically that for reaching consensus on a set of norms. We have defined an argumentation framework and a dialogue protocol (in Chapter 8), which facilitates deliberation among coherence-driven normative agents by exchanging norm proposals and counter proposals. Unlike argumentation systems derived from Dung's argumentation framework, a coherence-driven argumentation offers flexibility in the notion of attack relations and set of accepted arguments.

Thus, with this framework, we have covered both cognitive and social aspects of autonomous behaviour of agents in sufficient depth and detail. We have not only provided both a philosophical and formal grounding for our framework (in Chapters 3, 4, 7 and 8), but also have provided an agent architecture with algorithmic specification of behaviour of coherence-driven agents in Chapter 5. Further, our simulation results on the game scenario provide sufficient proof that the proposed architecture is both computationally feasible and performs equally well, when compared to performances of human and near optimal algorithms. We, however, would like to note that, the current experimental set-up is still insufficient to show the specific advantages of a coherence-driven approach for agent design.

## 10.3 Theory of Coherence

This book centres around the notion of coherence and its applicability to the field of artificial agents and MAS. The motivation was always clear, to make artificial

agents more autonomous and to make societies of such agents to do activities normally not entrusted to agent systems. One can go through this book, fully acknowledging these motivations and making a complete read following how these motivations have been addressed and accomplished. Yet, we would like to bring to front another subtle thread that runs in parallel. This is the thread that treats the theory of coherence. As discussed in Chapter 2, Thagard's theory of coherence is a mixture of mathematical formulations, guiding principles, example scenarios and algorithmic specifications. This informal approach is due to the fact that Thagard primarily focused on making an explanatory tool, which for example, would aid in explaining, analysing and predicting human behaviour. Even so, it may be difficult for experimenters to make use of his tools as there are no precise functions which would ensure a unique coherence value between two pieces of information. Designing an autonomous agent centring around coherence was far from his field of applications. Thagard, while having made an important contribution defining coherence as maximising constraint satisfaction, at times drifted to the point of view of a cognitive scientist. A good example is the set of principles for computing types of coherence. We, from the stand of computer scientists and logicians, not only borrowed the theory for our use, but gave it a logical grounding, formalised, studied its properties as a logical relation, and finally provided a computational mechanism which enable computing coherence graphs from sets of pieces of information. This, not only will benefit the artificial intelligence community by trying to make use of coherence theory or similar theories, but also sociologists, psychologists, philosophers and even economists interested in the theory for purposes of their own. This is the contribution we offer to the social sciences, to the theory of coherence, a grounding of the theory in logic and a computational mechanism to compute coherence graphs given only pieces of information.

## 10.4 Future directions

A research work is interesting, not only when it provides valuable answers, but also when it generates a whole lot of new interesting questions. The interaction between artificial agents and theory of coherence has stimulated a whole new set of research directions. Since this book is spread into a number of research areas, there are a few research paths that is worth in exploring in all of these areas. We categorise them and present below those that are most significant.

### 10.4.1 Formalisation of Coherence

In this subsection, we elaborate those directions that are primarily related to the formalisation of the theory of coherence and the coherence framework we established in this book.

1. **Types of Coherence:** In this book, we used a specific type of coherence namely deductive coherence without knowing if it is the most prominent

relation that exists between pieces of information. However, when studying cognitive elements, some relationships are stronger than others and it is worth representing the most significant relation instead of trying to measure a single standard relation between all pairs of elements. For example, the strongest or most significant relation between a desire and an intention is *facilitation*, that an intention facilitating a desire. *e.g. the intention to eat facilitates a desire to eat.* According to Thagard [Thagard, 2002], this relation is characterised as *deliberative coherence*. Principles of deliberative coherence differ from that of deductive coherence in that, two intentions that independently facilitate a desire compete with each other or negatively cohere whereas two formulas used to deduce a third formula consistent or do not negatively cohere in deductive coherence.

The choice of deductive coherence in this book is deliberate and is influenced by the well defined nature of logical deduction, whereas theories of explanation or facilitation lack such well defined theories, are precise and can be readily used. The future work may aim to characterise different types of coherence following the formalisation of deductive coherence of this book and a coherence graph may be defined in such a way as to represent the most significant coherence relation between all pairs of elements in the graph. In the context of artificial agents, one of the appropriate coherence relation that might be studied is deliberative coherence and particular emphasis may be given to its exploration.

2. **Introducing a Context:** In the current work, coherence maximisation is performed considering the entire set of cognitive elements. However, in reality, an action is chosen not just due to the fact that the said action is part of an accepted set, but, more because the said action is in an accepted set along with certain specific beliefs and goals. Thus, there is a requirement of an intuitive context that is necessary for an action to be a right action. We define this as the context for action and it would be fruitful to make precise this intuitive notion of context in the coherence framework and associated computations.

### 10.4.2 Coherence and Autonomous Agents

This book stands to support the use of coherence in designing autonomous artificial agents. However, the present work leaves many interesting associations unexplored. A number of surrounding research areas that may benefit from a coherence-based interpretation are the following:

1. **Cognitive Revision:** In this book, we mention that the process of coherence maximisation is similar to the process of theory revision. The process of coherence maximisation in effect drives a process of theory revision. The intuition is that, if a belief in the accepted set moves to the rejected set as a result of a coherence maximisation, the confidence on the belief no longer remains the same and should be reduced. The contrary is true if a rejected

belief is accepted again. Thus, from an initial approximate assignment of grades to cognitive elements, each time a coherence maximisation moves a cognitive element from accepted to rejected set, a decrease in grade can be computed using a probability measure. Thus, over a number of revisions, a grade associated with a cognitive element may tend towards the true grade (confidence, priority). This, if developed, might solve one of the important bottlenecks in realising cognitive agents with a representation of uncertainty. This is one of the important future works and needs to be explored further. Coherence maximisation as a cognitive revision mechanism also may be analysed in the context of AGM theories for belief revision [Koons, 2009].

### 10.4.3 Coherence and Normative MAS

In this book, we have embarked on a new agenda for regulated MAS to be adaptive. To be justified, we have only developed the necessary building blocks for norm adaptation. Here, we explore some of the most relevant future work in the context of MAS.

1. **Interaction of Agents in Normative Systems :** The evolution of a regulated system has many phases to be taken into consideration. We have touched upon one of the phases, namely norm acceptance. Norm acceptance may come into play when a regulatory system is being set up or when the need to change some of the norms becomes evident. In both cases, agents need to agree on a new set of norms through deliberation or other agreement generation techniques. However, followed by the norm acceptance phase, the agreed upon norms need to be established in the community by enacting the enforcement. For norm adoption to happen, agents need to have a representation of norms, should recognise the presence of a norm and understand the consequences of it. For these to be performed efficiently, a sufficiently expressive representation of norms is essential. In this book, we have a simple representation of norms, which is not developed enough for the mentioned purposes. One of the essential future works is to develop an expressive representation of norms considering its interaction within a regulated MAS.

   Modelling the regulated environment is one of the essential parts of modelling the interactions of normative agents with their environment. A prominent line of work uses temporal logic for normative systems [Ågotnes et al., 2007]. Norms are interpreted as constraints on potential agent behaviours and hence are interpreted as forbidden state transitions. The introduction of temporal constraints makes this normative system very expressive. Any normative action an autonomous agent takes should be reflected in such a normative environment, and an interesting future work is to explore this coupling between a normative environment and autonomous normative agents.

2. **Performance Criteria for Argumentation:** In this book, we have used a coherence-based argumentation system for deliberation on norm adoption. In general argumentation is seen as a viable means for multi agent deliberation (negotiation, persuasion) in general and for deliberation on norms in particular. However, in the argumentation literature, there are no established performance criteria for measuring the quality of an argument in terms of the positions of the participating agents or in terms of the common goals they set out to achieve. Such measures are very desirable and even essential because they are indicative of the quality of decisions made by a MAS and hence can be used to determine both the efficiency and the reliability. More importantly, these measures can guide agents towards an equilibrium outcome. The MAS and argumentation research communities have started identifying the need for such criteria for generic agents using argumentation [Rahwan et al., 2009], focusing however on strategy-proofness properties of very specific argumentation semantics.

   A need for performance measures is also present in the case of coherence-driven argumentation. The only well known measures for strategic interaction among rational agents are Nash equilibrium and Pareto optimality [Fudenberg and Tirole, 1991]. They are however defined under strict notions of rationality and assumptions of perfect information. Even with their limitations, a possible benefit may be that these properties are well established and have rigourous mathematical proofs. An interesting future work may be to study how notions of Nash equilibrium and Pareto optimality from the game theory literature can be adapted to define performance criteria for argumentative coherence-driven agents.

### 10.4.4 Applications

We conclude the future directions by giving a feel of the kind of applications that might benefit from a coherence-driven approach. One class of such applications is automated dispute resolution where the participants have conflicting interests. Here we apply coherence-driven agents in one such case of a real political conflict.

We show how a cognitive agent endowed with a coherence-based architecture is capable of taking decisions by means of maximising the coherence of its beliefs, desires and intentions. The example is motivated by the water sharing treaty signed between the southern states of India during 1892 and 1924 and the disputes thereafter [Wikipedia, 2008]. We simplify the case for brevity: we model the reasoning of just one of the agents (southern Indian state $s$) involved in the conflict in three snapshots of time 1891, 1892, and 1991, the first one when the first treaty is about to be signed (when the decision to adopt a norm is to be taken), the second, when the norm is adopted and the third after a long period of co-operation between the states, when the situation had significantly evolved and the norm is to be broken by $s$.

### 10.4.5 Terminology

To represent the cognitions and norms of an agent, we shall use belief, desire, intention and norm languages as defined in Section 5.2.1. Hence, $(B\varphi, r)$ represents that the agent believes that proposition $\varphi$ is true with degree at least $r$. (Propositions $(D\varphi, r)$, and $(I\varphi, r)$ are desires and intentions and are interpreted analogously.) $(O\varphi, r)$ is the obligation of the agent to make $\varphi$ happen. The degree $r$ is a measure on the relevance of the norm, such as for instance its priority, or to what extend it needs to be fulfilled. The statements about the world are in a propositional language where each proposition is a grounded predicate with obvious meaning as can be seen later in the snapshots.

We assume that the agent has four contexts $C_{\mathrm{B}}$, $C_{\mathrm{D}}$,$C_{\mathrm{I}}$, and $C_{\mathrm{N}}$ containing the beliefs, desires, intentions, and permissions/obligations, respectively. We will omit the reference to the contexts in the notation as all beliefs are in $C_{\mathrm{B}}$, desires in $C_{\mathrm{D}}$, intentions in $C_{\mathrm{I}}$, and permissions and obligations in $C_{\mathrm{N}}$, and therefore there is no possible confusion. The two bridge rules we use in the water-sharing example are the following:

- $b_1 = \dfrac{(B\varphi, r) \quad (D\varphi, s)}{(I\varphi, \min(r, s))}$: Whenever a proposition is believed with degree at least $r$ and desired with degree at least $s$, then a corresponding intention with a degree at least $\min(r, s)$ is added to the theory of context $C_{\mathrm{I}}I$. We don't intend stronger than we desire or we believe.

- $b_2 = \dfrac{(B\varphi, r) \quad (O\varphi, s)}{(I\varphi, \min(r, s))}$: If the agent beliefs that an obligation is feasible, then it intends to make it happen.

### 10.4.6 Pre-treaty situation (1891)

The following tables and graph represent the situation before the treaty between the two states is proposed.

| | |
|---|---|
| $\mathcal{T}_B$ | $\{(B\varphi_1, 0.75), (B\varphi_2, 0.9), (B\varphi_5, 1)\}$ |
| $\mathcal{T}_D$ | $\{(D\varphi_3, 0.95)\}$ |
| $\mathcal{T}_B^{\bullet}$ | $\mathcal{T}_B \cup \{(B\varphi_4, 0.68), (B\varphi_3, 0.68)\}$ |

Table 10.1: The theories of $s$ and the deductive closure of context $B$. In this case only the belief context deduces new formulas.

In Table 10.1, we list the elements of the theories $\mathcal{T}_B, \mathcal{T}_D$ and the subformula-closure of context $C_{\mathrm{B}}$ for agent $s$ before the norm was proposed. The propositions $\varphi_i$ are as given in Table 10.2. The coherence graph, $g_1$ obtained from these theories at the end of Step 5 of the agent reasoning process (see Section 5.3) is represented in Figure 10.1, that is, including the effects of the deductions by means of bridge rules (i.e. $(I\varphi_3, 0.68)$). The coherence $\kappa(g_1)$ is 0.32 and the accepted set $\mathcal{A}$ includes all the nodes in the graph. The computation

| | |
|---|---|
| $\varphi_1$ | $good(rainfall)$ |
| $\varphi_2$ | $adequate(waterlevel)$ |
| $\varphi_3$ | $satisfied(demand)$ |
| $\varphi_4$ | $\varphi_1 \wedge \varphi_2$ |
| $\varphi_5$ | $\varphi_4 \rightarrow \varphi_3$ |

Table 10.2: Propositions relevant for the cognitions of $s$ at the beginning of the reasoning.

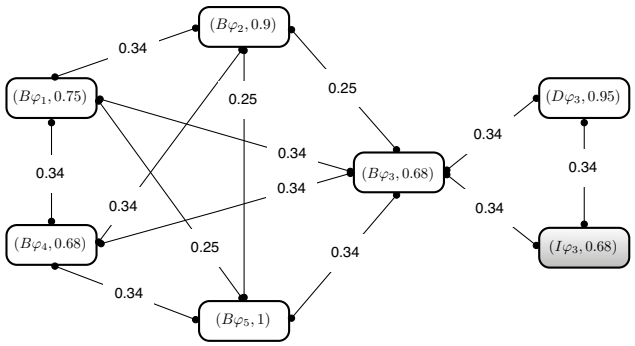


Figure 10.1: Initial coherence graph ($g_1$) of $s$ as in 1891 including the bridge rule deductions (shadowed) with $\kappa(g_1) = 0.32$

is illustrated in one case: $\eta((B\varphi_1, 0.75), (B\varphi_2, 0.9)) = \frac{0.65}{2} = 0.38$, given that $(B\varphi_1, 0.75), (B\varphi_2, 0.9) \vdash (B\varphi_4, 0.75 * 0.9)$ (assuming probabilistic independence) where $\Gamma$ is empty, and there is no other possible deduction to obtain one of the formulas from the other.

**Norm adoption. Evaluating the Treaty (1892).**

In 1892, a new norm was proposed: the Indian state $s$ will get obliged to release 300 billion ft$^3$ of water to its neighbour state annually; included in the proposal there was a threat of military retaliation in the case of unfulfillment of the obligation. Certainly, the release of water might threaten the objective of satisfying the internal demand and state $s$ was not necessarily happy with it. The situation of the theories at the beginning of the treaty is as expressed in Table 10.3. We have added the new formulas associated with the obligation and its related facts.

Agent $s$ evaluates the proposal of the new treaty by incorporating into its theories and its respective coherence graphs the new obligation, its implications and the sanctions that might be incurred if the proposal is not accepted. That is, the theories are updated according to Table 10.3, where the relevant propositions $\varphi_i$ are as in Table 10.4.

| *Theory* | *Existing* | *New* |
|---|---|---|
| $\mathcal{T}_N$ | | $\{(O\varphi_6, 1)\}$ |
| $\mathcal{T}_B$ | $\{(B\varphi_1, 0.75), (B\varphi_2, 0.9), (B\varphi_5, 1),$ $(B\varphi_4, 0.75), (B\varphi_3, 0.75)\}$ | $\{(B\varphi_{10}, 0.85),$ $(B\varphi_9, 0.9), (B\varphi_7, 0.7)\}$ |
| $\mathcal{T}_D$ | $\{(D\varphi_3, 0.95)\}$ | $\{(D\neg\varphi_7, 1)\}$ |
| $\mathcal{T}_I$ | $\{(I\varphi_3, 0.75)\}$ | |

Table 10.3: New elements introduced into the theory of $s$ in 1892

| | |
|---|---|
| $\varphi_6$ | $release(300\ billion\ ft^3)$ |
| $\varphi_7$ | $realised(attack)$ |
| $\varphi_8$ | $\varphi_1 \wedge \varphi_2 \wedge \varphi_6$ |
| $\varphi_9$ | $\varphi_8 \rightarrow \neg\varphi_3$ |
| $\varphi_{10}$ | $\neg\varphi_6 \rightarrow \varphi_7$ |

Table 10.4: Propositions relevant for the cognitions of $s$ in 1892

Agent $s$ now computes the composite coherence graph $g_2$ (shown in Figure 10.2) resulting from the theory update and using the set of bridge rules $B = \{b_1, b_2\}$. There are no negative coherence values between any pair of cognitions so the whole set is accepted again. However, this time the overall strength of the maximal partition is $\kappa(g)$ is 0.225. It is clear that coherence has decreased by incorporating the new norm which might be interpreted as an indication that the overall situation for $s$ was not as good as before signing the treaty. But still the accepted set includes the acceptance of the norm. Hence, guided by coherence maximisation, agent $s$ signs the treaty.

### 10.4.7 The Incoherence Buildup (1991)

| *Theory* | *Existing* | *New* |
|---|---|---|
| $\mathcal{T}_N$ | $\{(O\varphi_6, 1)\}$ | |
| $\mathcal{T}_B$ | $\{(B\varphi_1, 0.75), (B\varphi_2, 0.9), (B\varphi_3, 0.75),$ $(B\neg\varphi_3, 0.15), (B\varphi_4, 0.75), (B\varphi_5, 1),$ $(B\varphi_6, 0.26), (B\varphi_7, 0.7), (B\varphi_8, 0.17),$ $(B\varphi_9, 0.9), (B\varphi_{10}, 0.85)\}$ | $\{(B\varphi_{12}, 0.9),$ $(B(\varphi_{12} \rightarrow \varphi_{11}), 0.8)$ $(B(\varphi_{11} \rightarrow \neg\varphi_3), 0.8)\}$ |
| $\mathcal{T}_D$ | $\{(D\varphi_3, 0.95), (D\neg\varphi_7, 1)\}$ | $\{(D\varphi_{12}, 0.85)\}$ |
| $\mathcal{T}_I$ | $\{(I\varphi_3, 0.68), (I\varphi_6, 0.26), (I\neg\varphi_7, 0.7)\}$ | |

Table 10.5: New elements introduced into the theories of $s$ in 1991

By 1991 $s$ experiences large-scale industrialisation, urbanisation, and higher revenue growth and as a consequence $s$ also experiences higher water usage. Specially important for the example is the fact that an increase in water usage means that the possibility of satisfying the internal demand will decrease as the fact $(B(\varphi_{11} \rightarrow \neg\varphi_3), 0.8)$ indicates (see Table 10.5).

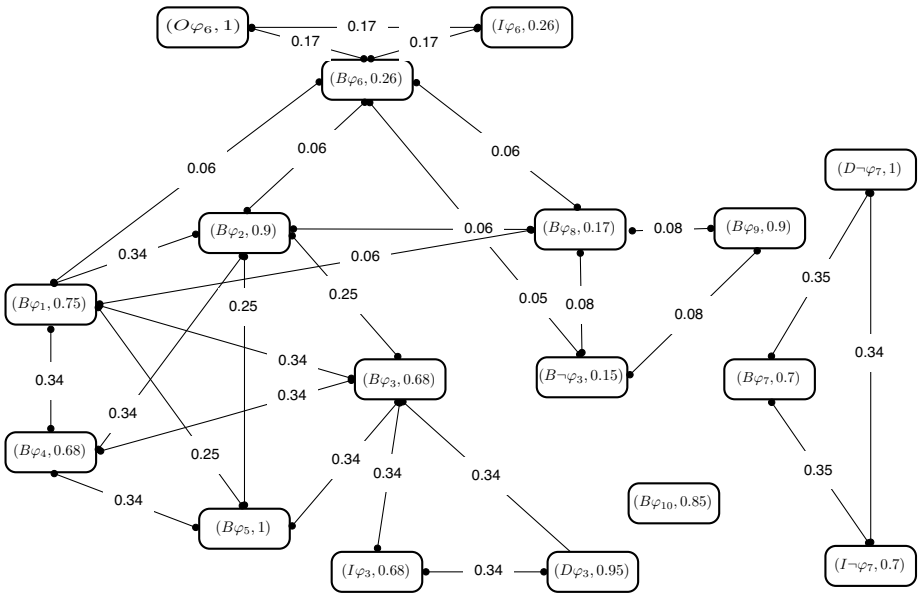


Figure 10.2: Coherence graph ($g_2$), with norm accepted $\kappa(g_1) = 0.225$

| | |
|---|---|
| $\varphi_{11}$ | $increase(waterusage)$ |
| $\varphi_{12}$ | $higher(growth)$ |
| $\varphi_{13}$ | $(B\varphi_{12} \rightarrow \varphi_{11}, 0.8)$ |
| $\varphi_{14}$ | $(B\varphi_{11} \rightarrow \neg\varphi_3, 0.8)$ |

Table 10.6: Propositions relevant for the cognitions of $s_1$ in 1991

In this last time point we don't depict the complete graph (see Figure 10.3) as it is too large, and we just show some of the relevant nodes where the incoherence has built up. The facts involved are listed in Table 10.6.

The incoherence builds up as the degree of belief in $\neg\varphi_3$ increases as a consequence of the new added facts.[1] That creates a negative coherence that forces for the first time a partition with maximum coherence that does not include all the nodes in the graph. Using the criterion of selecting as the accepted set of the partition the subgraph with maximal strength the agent chooses the set on the right as the accepted one. In particular, node $(O\varphi_6, 1)$ appears in the rejected set and therefore the Indian state $s$ decides to break the norm in order to keep a maximal coherence value.

As we have seen, such a reasoning takes into account the effects of norms in the cognitive state of an agent. As described in Chapter 8, a decision to violate

[1] We assume the inference rule $(B\varphi, r), (B\neg\varphi, s) \vdash (\bar{0}, 1 - (r + s))$.

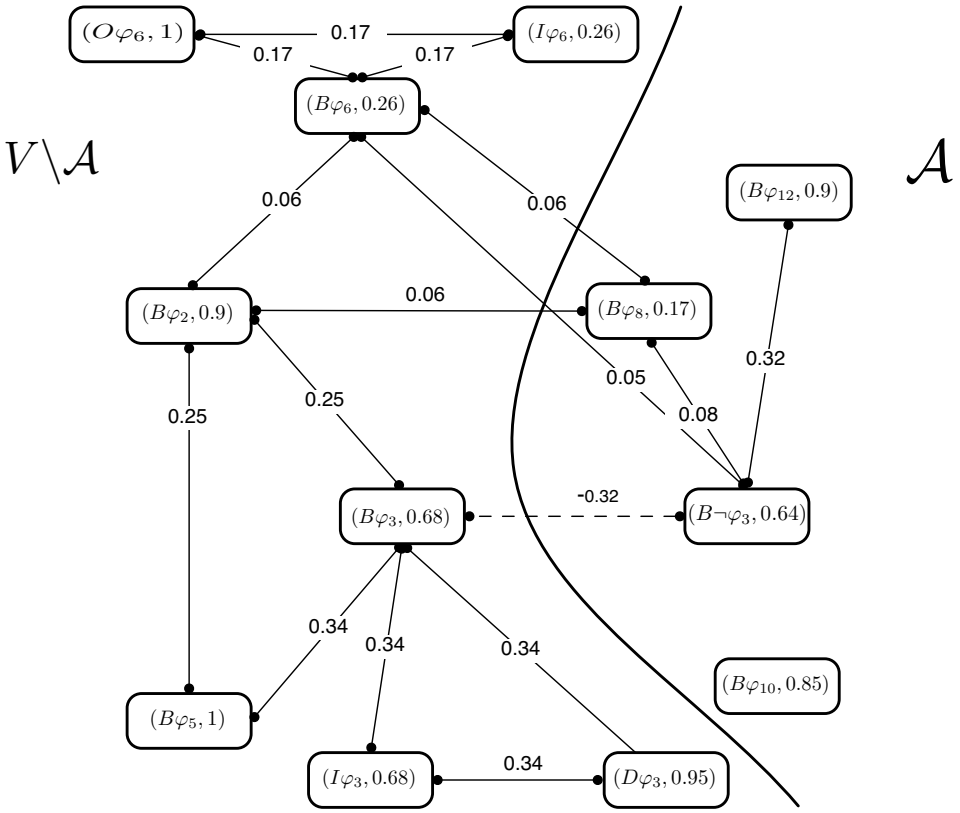


Figure 10.3: Subgraph of the coherence graph ($g_3$)

a norm may trigger deliberations that may lead to a re-definition of the norm and subsequently, an adaptation of the corresponding normative MAS to better fit situational changes.

# Bibliography

[Ågotnes et al., 2007] Ågotnes, T., Van Der Hoek, W., Rodríguez-Aguilar, J. A., Sierra, C., and Wooldridge, M. (2007). On the logic of normative systems. In *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1175–1180, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Amaya, 2007] Amaya, A. (2007). Formal models of coherence and legal epistemology. *Artificial intelligence and law*, 15(4):429–447.

[Amaya, 2009] Amaya, A. (2009). Inference to the best legal explanation. In Kaptein, H., Prakken, H., and Verheij, B., editors, *Legal Evidence and Proof: Statistics, Stories, Logic*. Ashgate Publishing, Aldershot.

[Amgoud et al., 2008] Amgoud, L., Cayrol, C., Lagasquie-Schiex, M. C., and Livet, P. (2008). On bipolarity in argumentation frameworks. *Int. J. Intell. Syst.*, 23(10):1062–1093.

[Amgoud et al., 2000] Amgoud, L., Maudet, N., and Parsons, S. (2000). Modeling dialogues using argumentation. In *ICMAS '00: Proceedings of the Fourth International Conference on MultiAgent Systems*. IEEE Computer Society.

[Amgoud and Prade, 2009] Amgoud, L. and Prade, H. (2009). Using arguments for making and explaining decisions. *Artificial Intelligence*, 34:197–216.

[Aranda et al., 2008] Aranda, G., Carrascosa, C., and Botti, V. (2008). Characterizing massively multiplayer online games as multi-agent systems. In *HAIS '08: Proceedings of the 3rd international workshop on Hybrid Artificial Intelligence Systems*, pages 507–514. Springer-Verlag.

[Atkinson, 2005a] Atkinson, K. (2005a). *What Should We Do?: Computational Representation of Persuasive Argument in Practical Reasoning.* PhD thesis, University of Liverpool.

[Atkinson, 2005b] Atkinson, K. (2005b). *What Should We Do?: Computational Representation of Persuasive Argument in Practical Reasoning.* PhD Thesis, Department of Computer Science, University of Liverpool, Liverpool, UK.

[Avron, 1991] Avron, A. (1991). Simple consequence relations. *Inf. Comput.*, 92(1).

[Bench-Capon and Prakken, 2006] Bench-Capon, T. and Prakken, H. (2006). Justifying actions by accruing arguments. In Dunne, P. and Bench-Capon, T., editors, *Computational Models of Argument. Proceedings of COMMA 2006*, pages 247–258, Amsterdam etc. IOS Press.

[Bench-Capon and Sartor, 2001] Bench-Capon, T. and Sartor, G. (2001). A quantitative approach to theory coherence. In Verheij, B., Lodder, A., Loui, R., and Muntjewerff, A., editors, *Legal Knowledge and Information Systems. JURIX 2001: The Fourteenth Annual Conference*, pages 53–62, Amsterdam etc. IOS Press.

[Blackburn et al., 2006] Blackburn, P., Benthem, J. F. A. K. v., and Wolter, F. (2006). *Handbook of Modal Logic, Volume 3 (Studies in Logic and Practical Reasoning)*. Elsevier Science Inc., New York, NY, USA.

[Boella et al., 2009] Boella, G., Caire, P., and van der Torre, L. (2009). Norm negotiation in online multi-player games. *Knowl. Inf. Syst.*, 18(2):137–156.

[Boella et al., 2006] Boella, G., Torre, L., and Verhagen, H. (2006). Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, 12(2-3).

[Boella and van der Torre, 2004] Boella, G. and van der Torre, L. (2004). Fulfilling or violating obligations in normative multiagent systems. In *IEEE/WIC/ACM International Conference on Intelligent Agent Technology*.

[Boella et al., 2007] Boella, G., van der Torre, L. W. N., and Verhagen, H., editors (2007). *Normative Multi-agent Systems, 18.03. - 23.03.2007*, volume 07122 of *Dagstuhl Seminar Proceedings*. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

[Bratman, 1987] Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. CSLI publications.

[Bratman et al., 1988] Bratman, M. E., Israel, D. J., and Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355.

[Broersen et al., 2001] Broersen, J., Dastani, M., Hulstijn, J., Huang, Z., and van der Torre, L. (2001). The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *AGENTS '01: Proceedings of the fifth international conference on Autonomous agents*, pages 9–16. ACM.

[Broersen et al., 2002] Broersen, J., Dastani, M., Hulstijn, J., and van der Torre, L. (2002). Goal generation in the BOID architecture. *Cognitive Science Quarterly Journal*, 2:428–447.

[Casali et al., 2005] Casali, A., Godo, L., and Sierra, C. (2005). Graded BDI models for agent architectures. In *Computational Logic in Multi-Agent Systems (CLIMA V)*, volume 3487 of *LNAI*, pages 126–143, Berlin/Heidelberg. Springer.

[Casali et al., 2006] Casali, A., Godo, L., and Sierra, C. (2006). A methodology to engineer graded BDI agents. In *WASI - CACIC Workshop.XII Congreso Argentino de Ciencias de la Computaci?n.*

[Castelfranchi et al., 2000] Castelfranchi, C., Dignum, F., Jonker, C. M., and Treur, J. (2000). Deliberative normative agents: principles and architecture. In *ATAL '99: 6th International Workshop on Intelligent Agents VI, Agent Theories, Architectures, and Languages*, pages 364–378. Springer-Verlag.

[Cesta et al., 2003] Cesta, A., Bahadori, S., Cortellessa, G., Grisetti, G., Giuliani, M. V., Iocchi, L., Leone, G. R., Nardi, D., Oddi, A., Pecora, F., Rasconi, R., Saggese, A., and Scopelliti, M. (2003). The robocare project cognitive systems for the care of the elderly. In *In Proceedings of International Conference on Aging, Disability and Independence (ICADI.*

[Conte, 2001] Conte, R. (2001). Emergent (info)institutions. *Cognitive Systems Research*, 2:97–110.

[Conte et al., 1999] Conte, R., Castelfranchi, C., and Dignum, F. (1999). Autonomous norm acceptance. In *ATAL '98: Proceedings of the 5th International Workshop on Intelligent Agents V, Agent Theories, Architectures, and Languages*, volume 1555 of *LNCS*, pages 99–112, Berlin/Heidelberg. Springer.

[Dastani et al., 2003] Dastani, M., de Boer, F., Dignum, F., and Meyer, J.-J. (2003). Programming agent deliberation: an approach illustrated using the 3apl language. In *AAMAS '03*, pages 97–104. ACM.

[Dellunde and Godo, 2008] Dellunde, P. and Godo, L. (2008). Introducing grades in deontic logics. In *DEON '08: Proceedings of the 9th international conference on Deontic Logic in Computer Science*, volume 5076 of *LNAI*, pages 248–262, berlin/Heidelberg. Springer.

[Dennett, 1971] Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy*, 68(February):87–106.

[Dung, 1995] Dung, P. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and *n*–person games. *Artificial Intelligence*, 77:321–357.

[Dunne and Bench-Capon, 2002] Dunne, P. E. and Bench-Capon, T. J. M. (2002). Coherence in finite argument systems. *Artificial Intelligence*, 141(1):187–203.

[Festinger, 1957] Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford University Press.

[Fitoussi and Tennenholtz, 2000] Fitoussi, D. and Tennenholtz, M. (2000). Choosing social laws for multi-agent systems: minimality and simplicity. *Artificial Intelligence*, 119(1-2):61–101.

[Fudenberg and Tirole, 1991] Fudenberg, D. and Tirole, J. (1991). *Game Theory*. MIT Press.

[Garson, 2009] Garson, J. (2009). Modal logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab,Stanford University, fall 2009 edition.

[Giunchiglia and Giunchiglia, 1993] Giunchiglia, F. and Giunchiglia, F. (1993). Contextual reasoning. *Epistemologia, special issue on I Linguaggi e le Macchine*, 345:345–364.

[Giunchiglia and Serafini, 1994] Giunchiglia, F. and Serafini, L. (1994). Multilanguage hierarchical logics, or: how we can do without modal logics. *Artificial Intelligence*, 65:29–70.

[Gordon, 1994] Gordon, T. (1994). The Pleadings Game: an exercise in computational dialectics. *Artificial Intelligence and Law*, 2:239–292.

[Hájek, 1998] Hájek, P. (1998). Metamathematics of fuzzy logic. *Trends in Logic*, 4(February).

[Hájek, 1998] Hájek, P. (1998). Metamathematics of fuzzy logic. In *Trends in Logic*, volume 4.

[Hamblin, 1970] Hamblin, C. L. (1970). *Fallacies*. [London] Methuen.

[Joseph et al., 2008a] Joseph, S., Dellunde, P., Schorlemmer, W. M., and Sierra, C. (2008a). Formalizing deductive coherence: An application to norm evaluation. In *NORMAS*, pages 158–172.

[Joseph and Prakken, 2009] Joseph, S. and Prakken, H. (2009). Coherence-driven argumentation to norm consensus. In *ICAIL '09: Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 58–67. ACM.

[Joseph et al., 2009a] Joseph, S., Schorlemmer, M., and Sierra, C. (2009a). Coherence as an inclusive notion of rationality. In *Proceeding of the 2009 conference on Artificial Intelligence Research and Development:Proceedings of the 12th International Conference of the Catalan Association for Artificial Intelligence*, pages 224–233. IOS Press.

[Joseph et al., 2008b] Joseph, S., Sierra, C., and Schorlemmer, M. (2008b). A coherence based framework for institutional agents. In *Coordination, Organizations, Institutions, and Norms in Agent Systems III*, volume 4870 of *LNCS*, pages 287–300, Berlin/Heidelberg. Springer.

[Joseph et al., 2010] Joseph, S., Sierra, C., and Schorlemmer, M. (2010). Cognitive coherence driven action selection in dynamic environments (extended abstract). In *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*.

[Joseph et al., 2009b] Joseph, S., Sierra, C., Schorlemmer, M., and Dellunde, P. (2009b). Deductive coherence and norm adoption. *Logic Journal of IGPL*, doi:10.1093/jigpal/jzp074.

[Kollingbaum and Norman, 2003] Kollingbaum, M. J. and Norman, T. J. (2003). Norm adoption in the noa agent architecture. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 1038–1039.

[Koons, 2009] Koons, R. (2009). Defeasible reasoning. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Stanford University, spring 2009 edition.

[K.Popper, 1962] K.Popper (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Basic Books.

[Kuhn, 1962] Kuhn, T. S. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Chicago: University of Chicago Press.

[Laird et al., 1991] Laird, J., Hucka, M., Huffman, S., and Rosenbloom, P. (1991). An analysis of soar as an integrated architecture. *SIGART Bull.*, 2(4):98–103.

[Laird et al., 1987] Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). Soar: an architecture for general intelligence. *Artif. Intell.*, 33(1):1–64.

[Lakatos, 1976] Lakatos, I. (1976). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Cambridge University Press.

[López et al., 2002] López, y. F. L., Luck, M., and d'Inverno, M. (2002). Constraining autonomy through norms. In *First International Joint Conference on Autonomous Agents and Multiagent Systems*.

[Loui, 1998] Loui, R. (1998). Process and policy: resource-bounded non-demonstrative reasoning. *Computational Intelligence*, 14:1–38.

[Luck et al., 2005] Luck, M., McBurney, P., Shehory, O., and Willmott, S. (2005). *Agent Technology: Computing as Interaction (A Roadmap for Agent Based Computing)*. AgentLink.

[Maes, 1991] Maes, P. (1991). *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*. MIT Press, Cambridge, MA, USA.

[Modgil, 2008] Modgil, S. (2008). An argumentation based semantics for agent reasoning. In *Languages, Methodologies and Development Tools for Multi-Agent Systems: First International Workshop, LADS 2007, Durham, UK, September 4-6, 2007. Revised Selected Papers*, pages 37–53. Springer-Verlag.

[Mook, 1987] Mook, D. G. (1987). *Motivation: The Organization of Action*. W. W. Norton & Company, New York.

[Moses and Tennenholtz, 1995] Moses, Y. and Tennenholtz, M. (1995). Artificial social systems. *Computers and AI*, 14:533–562.

[Noriega, 1997] Noriega, P. (1997). *Agent-Mediated Auctions: The Fishmarket Metaphor*. IIIA Phd Monography. Vol. 8.

[Paoli, 2002] Paoli, F. (2002). *Substructural Logics: A Primer*. Springer.

[Parsons et al., 1998] Parsons, S., Sierra, C., and Jennings, N. (1998). Agents that reason and negotiate by arguing. *JOURNAL OF LOGIC AND COMPUTATION*, 8:261–292.

[Pasquier et al., 2004] Pasquier, P., Andrillon, N., Labrie, M.-A., and Chaib-draa, B. (2004). An exploration in using cognitive coherence theory to automate BDI agents' communicational behavior. In *Advances in Agent Communication*. Springer.

[Pasquier and Chaib-draa, 2003] Pasquier, P. and Chaib-draa, B. (2003). The cognitive coherence approach for agent communication pragmatics. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 544–551. ACM.

[Pasquier et al., 2006] Pasquier, P., Rahwan, I., Dignum, F., and Sonenberg, L. (2006). Argumentation and persuasion in the cognitive coherence theory. In Dunne, P. and Bench-Capon, T., editors, *Computational Models of Argument. Proceedings of COMMA 2006*, pages 223–234, Amsterdam etc. IOS Press.

[Pitt, 2008] Pitt, D. (2008). Mental representation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Stanford University, fall 2008 edition.

[Piwek, 2007] Piwek, P. (2007). Meaning and dialogue coherence: a proof-theoretic investigation. *Journal of Logic, Language and Information*, 16(4):403–421.

[Pollock, 1975] Pollock, J. L. (1975). *Knowledge and justification*. Princeton University Press, New York, NY, USA.

[Prakken, 2005a] Prakken, H. (2005a). Coherence and flexibility in dialogue games for argumentation. *J. Log. and Comput.*, 15(6):1009–1040.

[Prakken, 2005b] Prakken, H. (2005b). A study of accrual of arguments, with applications to evidential reasoning. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law*, pages 85–94, New York. ACM Press.

[Prakken, 2009] Prakken, H. (2009). An abstract framework for argumentation with structured arguments. Technical Report UU-CS-2009-019, Department of Information and Computing Sciences, Utrecht University.

[Rahwan et al., 2009] Rahwan, I., Larson, K., and Tohmé, F. (2009). A characterisation of strategy-proofness for grounded argumentation semantics. In *IJCAI'09: Proceedings of the 21st international jont conference on Artifical intelligence*, pages 251–256. Morgan Kaufmann Publishers Inc.

[Rahwan et al., 2003a] Rahwan, I., Ramchurn, S., Jennings, N., McBurney, P., Parsons, S., and Sonenberg, L. (2003a). Argumentation-based negotiation. *The Knowledge Engineering Review*, 18:343–375.

[Rahwan et al., 2003b] Rahwan, I., Ramchurn, S. D., Jennings, N. R., McBurney, P., Parsons, S., and Sonenberg, L. (2003b). Argumentation-based negotiation. *The Knowledge Engineering Review*, 18:343–375.

[Rao and Georgeff, 1995] Rao, A. S. and Georgeff, M. (1995). BDI agents: From theory to practice. In *ICMAS–95, First International Conference on Multi-Agent Systems: Proceedings*, pages 312–319, S. Francisco, CA. MIT Press.

[Rao and Georgeff, 1991] Rao, A. S. and Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In Allen, J., Fikes, R., and Sandewall, E., editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

[Reiter, 1987] Reiter, R. (1987). A logic for default reasoning. *Readings in nonmonotonic reasoning*, pages 68–93.

[Russell and Norvig, 2003] Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education.

[Sansonnet and Valencia, 2003] Sansonnet, J.-P. and Valencia, E. (2003). A model for dialog between semantically heterogeneous informational agents. In *Eleventh Portuguese Conference on Artificial Intelligence*.

[Shoham, 1993] Shoham, Y. (1993). Agento: a simple agent language and its interpreter. In *Proceedings of AAAI*.

[Shoham and Tennenholtz, 1995] Shoham, Y. and Tennenholtz, M. (1995). On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73(1-2):231–252.

[Sierra et al., 2004] Sierra, C., Rodrguez-Aguilar, J. A., Noriega, P., Esteva, M., and Arcos, J. L. (2004). Engineering multi-agent systems as electronic institutions. *European Journal for the Informatics Professional*, 4.

[Simon, 1969] Simon, H. A. (1969). *The Sciences of the Artificial*. MIT Press.

[Sutton and Barto, 1998] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA.

[Thagard, 2002] Thagard, P. (2002). *Coherence in Thought and Action*. MIT Press.

[Thagard, 2004] Thagard, P. (2004). Causal inference in legal decision making: Explanatory coherence vs. Bayesian networks. *Applied Artificial Intelligence*, 18:231–249.

[Thagard, 2006] Thagard, P. (2006). *Hot Thought*. MIT Press.

[Walton, 1996] Walton, D. (1996). *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ.

[Walton and Krabbe, 1995] Walton, D. N. and Krabbe, E. C. W. (1995). *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, Albany.

[Wikipedia, 2008] Wikipedia (2008). Kaveri river water dispute — wikipedia, the free encyclopedia.

[Wooldridge, 2000] Wooldridge, M. (2000). *Reasoning about rational agents*. MIT press.

# Monografies de l'Institut d'Investigació en Intel·ligència Artificial

Num. 1 J. Puyol, *MILORD II: A Language for Knowledge–Based Systems*

Num. 2 J. Levy, *The Calculus of Refinements, a Formal Specification Model Based on Inclusions*

Num. 3 Ll. Vila, *On Temporal Representation and Reasoning in Knowledge–Based Systems*

Num. 4 M. Domingo, *An Expert System Architecture for Identification in Biology*

Num. 5 E. Armengol, *A Framework for Integrating Learning and Problem Solving*

Num. 6 J. Ll. Arcos, *The Noos Representation Language*

Num. 7 J. Larrosa, *Algorithms and Heuristics for Total and Partial Constraint Satisfaction*

Num. 8 P. Noriega, *Agent Mediated Auctions: The Fishmarket Metaphor*

Num. 9 F. Manyà, *Proof Procedures for Multiple-Valued Propositional Logics*

Num. 10 W. M. Schorlemmer, *On Specifying and Reasoning with Special Relations*

Num. 11 M. López-Sánchez, *Approaches to Map Generation by means of Collaborative Autonomous Robots*

Num. 12 D. Robertson, *Pragmatics in the Synthesis of Logic Programs*

Num. 13 P. Faratin, *Automated Service Negotiation between Autonomous Computational Agents*

Num. 14 J. A. Rodríguez, *On the Design and Construction of Agent-mediated Electronis Institutions*

Num. 15 T. Alsinet, *Logic Programming with Fuzzy Unification and Imprecise Constants: Possibilistic Semantics and Automated Deduction*

Num. 16 A. Zapico, *On Axiomatic Foundations for Qualitative Decision Theory - A Possibilistic Approach*

Num. 17 A. Valls, *ClusDM: A multiple criteria decision method for heterogeneous data sets*

Num. 18 D. Busquets, *A Multiagent Approach to Qualitative Navigation in Robotics*

Num. 19 M. Esteva, *Electronic Institutions: from specification to development*

Num. 20 J. Sabater, *Trust and Reputation for Agent Societies*
Num. 21 J. Cerquides, *Improving Algorithms for Learning Bayesian Network Classifiers*
Num. 22 M. Villaret, *On Some Variants of Second-Order Unification*
Num. 23 M. Gómez, *Open, Reusable and Configurable Multi-Agent Systems: A Knowledge Modelling Approach*
Num. 24 S. Ramchurn, *Multi-Agent Negotiation Using Trust and Persuasion*
Num. 25 S. Ontañon, *Ensemble Case-Based Learning for Multi-Agent Systems*
Num. 26 M. Sánchez, *Contributions to Search and Inference Algorithms for CSP and Weighted CSP*
Num. 27 C. Noguera, *Algebraic Study of Axiomatic Extensions of Triangular Norm Based Fuzzy Logics*
Num. 28 E. Marchioni, *Functional Definability Issues in Logics Based on Triangular Norms*
Num. 29 M. Grachten, *Expressivity-Aware Tempo Transformations of Music Performances Using Case Based Reasoning*
Num. 30 I. Brito, *Distributed Constraint Satisfaction*
Num. 31 E. Altamirano, *On Non-clausal Horn-like Satisfiability Problems*
Num. 32 A. Giovannucci, *Computationally Manageable Combinatorial Auctions for Supply Chain Automation*
Num. 33 R. Ros, *Action Selection in Cooperative Robot Soccer using Case-Based Reasoning*
Num. 34 A. García-Cerdaña, *On some Implication-free Fragments of Substructural and Fuzzy Logics*
Num. 35 A. García-Camino, *Normative Regulation of Open Multi-agent Systems*
Num. 36 A. Ramisa Ayats, *Localization and Object Recognition for Mobile Robots*
Num. 37 C.G. Baccigalupo, *Poolcasting: an intelligent technique to customise music programmes for their audience*
Num. 38 J. Planes, *Design and Implementation of Exact MAX-SAT Solvers*
Num. 39 A. Bogdanovych, *Virtual Institutions*
Num. 40 J. Nin, *Contributions to Record Linkage for Disclosure Risk Assessment*
Num. 41 J. Argelich Romá, *Max-SAT Formalisms with Hard and Soft Constraints*
Num. 42 A. Casali, *On Intentional and Social Agents with Graded Attitudes*
Num. 43 A. Perreau de Pinnick Bas, *Decentralised Enforcement in Multiagent Networks*
Num. 44 I. Pinyol Catadau, *Milking the Reputation Cow: Argumentation, Reasoning and Cognitive Agents*
Num. 45 S. Joseph, *Coherence-based Computational Agency*