**Universitat Autònoma de Barcelona**
**Programa de Doctorat en Informàtica**

UAB
Universitat Autònoma de Barcelona

# Formalising Deductive Coherence: An Application to Norm Evaluation

October 14, 2008

Memòria presentada per **Sindhu Joseph**
per optar al Diploma d'Estudis Avançats
sota la direcció del **Prof. Carles Sierra** i
el **Dr. Marco Schorlemmer**.


**Institut d'Investigació en Intel·ligència Artificial**
**Consell Superior d'Investigacions Científiques**

# Preface

This dissertation is a contribution to the formalisation of Thagard's coherence theory. The term *coherence* is defined as the quality or the state of cohering, especially a logical, orderly, and aesthetically consistent relationship of parts. Cognitive coherence in particular is the coherence theory based explanation of the mind which evaluates the truth of a cognition in terms of it being a member of some suitably defined body of other cognitions: a body that is consistent, coherent, and possibly endowed with other virtues, provided these are not defined in terms of truth. Thus a coherent set is interdependent such that every element in it contributes to the coherence. We take Thagard's proposal of a coherence set as that of maximising satisfaction of constraints between elements and explore its use in normative multiagent systems. We demonstrate its use as a mechanism to introduce true autonomy in agents, particularly in a normative multiagent setting.

In particular, we propose a coherence-driven agent, an agent that is driven by coherence maximisation. To design such an agent, we introduce a general coherence framework with the necessary computable functions. Further we analyse the formal properties of coherence so that we could make the framework fully computational. We use a proof-theoretic characterisation of coherence based on the principles proposed by Thagard. For the purpose of demonstration, we focus on one particular type of coherence namely, deductive coherence. Our use of graded logic helps us to incorporate reasoning under uncertainty which is more realistic in the context of multiagent systems. Finally, we explore a scenario where a coherence-driven agent deliberates about norms in a multiagent system where there is competition for a common resource. We show how a coherence-driven agent decides to violate a norm guided by its coherence, while also considering other factors such as sanctions and rewards.

**Acknowledgements.** The research is partially supported by the OpenKnowledge[1] 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 Spanish project "Agreement Technologies" (CONSOLIDER-INGENIO 2010 CSD2007-0022)

[1] http://www.openk.org

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

A normative multiagent system is a multiagent system where the agent interactions are governed by norms. In these systems, norms are identified with obligations, permissions and prohibitions, which specify the ideal behaviour of agents. Such systems also consider constitutive norms to give a new meaning to certain behaviour of agents. While a normative multiagent system is prescriptive about agent behaviour, it does so within the framework of autonomous agents. That is, the system assumes agents to behave in an autonomous manner and reason about norms autonomously. This is because of the fact that the success of such systems does not depend on all the agents following the prescribed norms blindly, rather on having rational agents deliberating about the prescribed norms, evaluating their usefulness, selectively following those norms that improve their efficiency, and effecting a change when there are conflicts, inefficiencies, or situational changes that they can perceive.

From the perspective of an agent, norms are like a guide book for an agent to make sense of what goes on around and what is expected of it. Norms are often meant to have a positive influence in a normative multiagent system, however, there may be cases where their implementations fail to translate this. In certain other cases, agents may have beliefs or goals which are in conflict with some of the norms, or there may even be cases where norms are in conflict between themselves. Thus both from the perspectives of normative systems and that of individual agents, it is not beneficial to treat norms like a hardwired goal in the agent architecture, rather they should be treated like dynamic entities that are deliberated before being adopted or obeyed.

We are certainly not the first to identify this need, and there have been numerous attempts in the recent past explicitly addressing this issue [19, 9, 4, 20, 31, 17]. Many of these efforts are focused towards extending the cognitive agent theory (for instance the Belief, Desire, and Intention (BDI) theory) with explicit representation of norms such as in BOID [6], EMIL [9], and NoA [17], or propose a more comprehensive multiagent system architecture that is norm-aware as in [4]. However, apart from providing static-priority based autonomy[1] and recognising autonomous norm acceptance phases, a gap still exists to enrich agent theories with true autonomy.

[1] A norm priority agent will always prefer norm compliance over satisfaction of private goals when there is a conflict.

To enhance the autonomous capabilities of agents, we propose a normative agent theory which extends the BDI theory with the theory of coherence [26]. Coherence theory, when used to explain human reasoning, proposes that humans accept or reject a cognition (external or internal) depending on how much it contributes to maximise the constraints imposed by situations and other cognitions. Pasquier et al. [20] introduced the possibility of extending agent reasoning with Thagard's theory of coherence. While their contribution introduces the concept of coherence in the field of multiagent systems, they still do not clarify the nature of a coherence relation nor do they specify how a coherence graph can be constructed. Thus, a general treatment of coherence to be used to realise computational models is still called for.

According to the theory of coherence, there are coherence and incoherence relations between *pieces of information* depending on whether they support (yielding a positive constraint) or contradict (yielding a negative constraint) each other. If two pieces of information are not related, then, there is no coherence (constraint) between them. Due to the fact that coherence is evaluated based on constraints that exist between pairs of information, a graph representation is most intuitive. Normally a graph with nodes and weighted edges are used to represent the pieces of information and 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. Understanding these principles and deducing methods to compute the coherence values between them is fundamental to compute the overall coherence of a given coherence graph. Without this important formalisation, practical realisations of coherence are hard to imagine.

In this dissertation we have chosen to analyse one such type of coherence, namely deductive coherence, because the theorems of logical deduction from which it is derived are well understood. Our aim is to generate coherence values between pairs of information (in this case, formulas in a logical language) by formalising the relationship between coherence and logical entailment. Coherence as a logical relation is significant in itself and has important implications: it is tolerant to inconsistencies and allows us to work with deductive systems without certain structural rules such as weakening.

More specifically the research questions addressed in this dissertation are along two dimensions: the first is to look for an appropriate model or theory to extend the existing agent theories; the second is to check its computational validity. That is:

> How can we design normative agents with more autonomous capabilities such that they can rationally evaluate norms in the light of their cognitions?
>
> If the theory of coherence is proposed to extend agent theories for autonomy, is it computationally realisable? What are the tools required to make a fully computational model of a normative coherence-driven agent?

We address the first question by following other researchers [20] to propose the theory of coherence for autonomous normative agent design. Though the theory has been proposed earlier to extend BDI agents in the context of communication, it has not been proposed as a general theory in the context of normative systems. We address the

second question by proposing a clear method, illustrating how a coherence based agent can reason autonomously about norms and cognitions by the process of coherence maximisation. We list the important steps in the process as follows.
Given a set of graded agent cognitions and a set of graded norms (norms with priorities) of a normative multiagent system,

1. evaluate the coherence or incoherence relations between pairs of cognitions of the same type and between pairs of norms;

2. evaluate coherence or incoherence relations between pairs of cognitions of differing types and norms;

3. given that all possible coherence or incoherence relations are computed for cognitions and norms, evaluate the overall coherence of the agent as if the agent were to accept all the cognitions and all the norms;

4. separate the set of cognitions and norms into two sets by a process of coherence maximisation such that only elements of one set are considered to be accepted (considered valid);

5. and finally, based on specific agent characteristics, a coherence maximising action is pursued if the increase in coherence corresponding to the accepted set is substantially higher when compared to the case in which the agent pursues all cognitions and norms.

This dissertation is structured as follows. Chapter 2 discusses the state of the art by summarising some of the prominent works in related fields and by elaborating how our work stands in relation with them. In Chapter 3 we give a general introduction to the theory of coherence, which helps the reader to understand the basic notions of coherence and how it differs from other related theories. We then introduce Thagard's theory of coherence and contrast it with other decision theories. In Chapter 4 we introduce a generic coherence framework which can be used to create coherence-based 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. Chapter 5 narrows down the focus to one particular type of coherence, namely deductive coherence. In this chapter, we provide a proof theoretic formulation of deductive coherence which we use to build coherence graphs given a set of pieces of information. Formal properties of the deductive coherence function enable us to compute coherence values between pieces of information. With this work, we intent to demonstrate that our proposed framework is fully computational.

In Chapter 6, we define a coherence-driven agent as a cognitive agent whose aim is to maximise coherence. For this purpose we define certain specific graphs corresponding to a cognitive agent. We adapt concepts taken from multi-context systems so that our coherence-driven agent can reason across cognitions and norms. We later sketch a procedure an agent may follow in the context of a normative multiagent system. Chapter 7 gives a detailed example inspired from a real-world scenario where a few southern regions of India participate in a water sharing normative system to share a common commodity, water, according to needs and quantity available. We in particular

consider two representative regions with one releasing water and the other receiving it under the agreements of the treaty. We study the reasoning of one of the agents and demonstrate that how this coherence-driven agent autonomously makes choices to accept a norm based on coherence maximisation. Later, however, due to the changes in the situation, it discovers through coherence maximisation that it is impossible to keep its private goals and also follow the norm it has been committed to. This example opens the possibility of agents deliberating autonomously which brings new challenges and possibilities to the study of normative multiagent systems.

Finally in Chapter 8 we conclude the work by discussing the important issues addressed in this dissertation, the challenges left open, and a few pointers to some of the the ongoing and future work. In particular, we discuss two specific instances of the ongoing work. The first attempts to demonstrate that coherence is an holistic notion and is inclusive of rationality. For this purpose we show that, the utility maximisation concept from a game theoretic context can be reduced to a particular type of coherence graph. The second introduces a semantic interpretation of coherence inspired by Ruspini's degrees of consistency. This opens up new possibilities to give a computational representation of other types of coherence other than deductive. We conclude with bibliography of important references used for this dissertation.

# Chapter 2

# State of the Art

The objective of this dissertation is twofold: to help design increasingly sophisticated agents with autonomous capabilities, and to demonstrate how agents who have autonomous abilities would take flexible and dynamic decisions when faced with dynamic and uncertain scenarios. We are particularly interested to bring in autonomous agents in the context of normative systems and to demonstrate that from the point of view of an agent, autonomy helps in evaluating norms rather than following designer specifications without deliberation. On the other hand, we aim to formalise the theory of coherence in a generic and computationally plausible manner so as to build coherence-driven agents that are autonomous. To this end, in this chapter we present work done in the fields of autonomous agent deliberation, normative systems and autonomous norm evaluation, and previous approaches in formalising coherence.

## 2.1 Autonomous Agent Deliberation

From the years that agent theory came into existence, autonomy is one of the most desired features to be incorporated in the agent design. The first major step was made when a behaviour model of agents was proposed. The BDI model for artificial agents is based on the theory of rational action in humans put forward in 1988 by the philosopher M. Bratman [5]. BDI is fundamentally reliant on folk psychology which is the notion that our mental models of the world are theories. BDI logics are multi-modal logics developed by Rao and Georgeff during the 1990s. However, the BDI model of agents was an attempt to solve a problem that has more to do with planning than with the design of autonomous agents. Yet, the BDI model served as the base model on which others could build more sophisticated features. From BDICTL of Rao and Georgeff's, LORA (the logic of rational agents) [30] to BOID [6], there have been numerous proposals to incorporate various levels of autonomy in agent design. However, as mentioned in the introduction, other than incorporating certain static priority based reasoning components into agent theories, we still lack sophisticated reasoning tools to make agents autonomous entities.

The work of Pasquier et al. [20] is an attempt at bringing more autonomous and dy-

namic reasoning into agent theories. They propose a cognitive coherence based model of communication, argumentation and reasoning from an agents perspective. The authors have developed a computational model of cognitive coherence which could be used to extend the agents reasoning mechanism to include social commitments. Their work is based, like ours, on the characterisation of coherence as maximising constraint satisfaction proposed by Thagard [26]. Thagard in his characterisation of coherence, differentiates types of coherence that need to be accounted for in order to formalise coherence. In our proposal we develop further this idea of Thagard and take the first step in this direction by giving a proof-theoretic characterisation of deductive coherence. Our approach differs from Pasquier et al. because our research is centered on calculating coherence measures. We understand coherence as a tool not only for maintaining the cognitions of individual agents but also for that of agents' society. Thus, we are interested in developing the concept of coherence in an institutional setting.

## 2.2 Normative Systems and Autonomous Norm Deliberation

As described in the introduction, norms help agents to form certain behaviour expectations of their counterparts in a multiagent system, which in turn helps the system to work efficiently. In this sense normative systems provide a very promising model for multiagent interaction and co-ordination [4]. One of the early introductions of norms for multiagent co-ordination is the work on artificial social systems by Tennenholtz and colleagues [25, 19, 13]. The problem studied in artificial social systems is the design, emergence or more generally the creation of social laws. Shoham and Tennenholtz studied artificial social systems using notions of game theory. Continuing their work, there has been much research in normative multiagent systems both from the social and from the cognitive perspectives [8, 9, 31]. As our work mainly deals with the cognitive aspect of norms, the following discussion focuses on proposals from a cognitive perspective. We discuss two of the representative proposals below.

The work by Guido et al. [3] gives a comprehensive account of the situations faced by different types of agents in which they could possibly violate norms. Situations include: when there are contradictions between goals and obligations, when violation is preferred to possible sanction, when an agent is ignorant about a norm or consequences of it, or, when it is impossible to fulfil the obligation. This work also attempts to formalise some of these notions. What relates Guido et al.'s work and ours is that all these situations are somehow incoherent, and a coherence-driven agent can be used to model them. However their work does not address the reasoning within the agent.

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 how agents can acquire norms, how agents can violate norms, and how an agent can be autonomous [9, 22]. 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 once the agent has accepted this norm, deciding to conform to it. The first step according to the authors would form the normative belief, and the second step would create the normative goal or intention. Moving from normative belief to normative conformity would additionally need the existence of other private goals of the agent which would benefit from the normative goal. The work provides a set of rules for normative acceptance and conformity. The authors, though recognising the importance of norm acceptance, sidestep the problem of coming up with mechanisms for autonomous norm acceptance. That is, recognising a norm as a norm is not equivalent to evaluating the norm. 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 our work is complementary to theirs as we propose a mechanism for norm evaluation, which can be embedded in the process of norm emmergence proposed by the authors.

## 2.3 Formalising Coherence

Here we primarily analyse those proposals that formalise coherence. 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. However, there have been a few proposals in the field of linguistic coherence. Hence we take two representative samples and analyse them in more detail. Both these works concentrate on linguistic coherence which is the property of a text or conversation being semantically meaningful. However, from the formal perspectives, there are overlaps as the principles of coherence essentially stays the same. We compare and contrast their proposals and our work.

The work of Piwek in [21] attempts to model dialogue coherence in terms of generative systems based on natural deduction. The main argument in his work is that it is possible to generate coherent dialogues by relying on entailment relations in the agents knowledge base. The paper primarily deals with information seeking dialogue where the definition of whether an agent knows a fact is equated to whether can be logically entailed. This is an interesting way to look at dialogue coherence where the concern here 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 Valencia et al. [24] models agent dialogue based on the theory of dissonance. 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 further when this dissonance no longer persists to terminate the dialogue. It is curious to note that many authors who have used the theory of dissonance

in dialogue initiation and termination [20, 24] have not considered the fact that not all incoherences are dissonance. Further, dissonance seeks out specialised information or actions. The most important difference between the work of Valencia et al. and ours is that, for them coherence (or the lack of it) is a local phenomena concerning only the new arriving fact and the fact that it contradicts with, whereas for us coherence is a global phenomena affecting the entire knowledge base of the agent. As in the case of the previous work, the authors equate coherence with logical entailment.

## 2.4 Discussion

Apart from the above classification of the related work, there is a considerable amount of interest in incorporating coherence theories in the field of legal reasoning [1, 11]. In [1], Amaya tries to apply a notion of coherence in legal justification and studies the how notions of fairness and coherence are related. 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 to be used in conducting argument justification in legal reasoning. She has analysed different aspects of coherence and has studied 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 neither a formalisation herself. Another work on argumentation [11]apply a coherence based mechanism for practical reasoning systems. Thus, we see that coherence has a diverse audience. This gives us enough reasons to go ahead with a generic formalisation which could translate these requirements into practical realisations. Specially, the interest for coherence in the legal reasoning community confirms that coherence is a desired property of valid arguments and hence of reasoning.

# Chapter 3

# Theory of Coherence

Here we introduce the theory of coherence and provide a summary of Thagard's Theory of Coherence, which is the major inspiration and the base of this dissertation. We then interpret Thagard's theory as a decision theory and contrast it with other decision theories.

## 3.1 General Theory of Coherence

Some of the foundational questions in epistemology deal with the origin, structure, and nature of knowledge and justified belief. The regress problem is an important problem when studying the structure of how knowledge is acquired or belief is justified. One of the central questions in the regress problem is to know how one knows or is justified in believing some particular thing. Many epistemologists studying justification have attempted to argue for various types of chains of reasoning that can escape the regress problem.

1. The series is infinitely long, with every statement justified by some other statement.
2. The series forms a loop, so that each statement is ultimately involved in its own justification.
3. The series terminates with certain statements having to be self justifying.

There are two main schools of thought in answering this question, foundationalism and coherentism. The foundationlist reject answers 1 and 2 and argue that 3 is the valid answer. According to the foundationalist option, the series of beliefs terminates with special justified beliefs called basic beliefs: these beliefs do not owe their justification to any other beliefs from which they are inferred [12]. Coherentism, however, argues that the 2nd argument is the valid one.

Coherentism rejects the argument that the regress proceeds according to a pattern of linear justification. To avoid the charge of circularity, coherentists hold that an individual belief is justified circularly by the way it fits together (coheres) with the rest

of the belief system of which it is a part. This theory has the advantage of avoiding the infinite regress without claiming special, possibly arbitrary status for some particular class of beliefs. There is nothing within the definition of coherence which makes it impossible for two entirely different sets of beliefs to be internally coherent. Thus, there might be several such sets, and pure coherentism does not offer a solution. However later theories of coherence admits certain favorable statements whose presence in a set makes it more coherent than other competing sets. These special statements are some of the obvious statements (which does not need justification). This sometimes is described as the meeting point between foundationalism and coherentism [18].

Even if one rejects the pure theory of coherence, one cannot deny the fact that, the property of coherence is a *necessary*, if not a *sufficient*, property of a system of justified beliefs or knowledge. This view on coherence has given raise to many applications of the theory in the field of philosophy and psychology. Recently, computer scientists have been increasingly taking a look at coherence and their applications in modelling behaviour of artificial entities such as agents. Though, the theory of coherence has been around for long, it was only recently, when the philosopher scientist Paul Thagard proposed a model of coherence as maximisation of constraint satisfaction, that the abstract theory of coherence became conceivable and even computable. Because this dissertation bases its foundations on this theory, we introduce it here.

## 3.2 Thagard's Theory of Coherence

Thagard postulates that coherence theory 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 [26, 27]. Thagard takes the theory from its abstract form and gives concrete interpretations of it. More importantly Thagard attempts to extend the theories reach to a broad audience by explaining how the theory of constraint satisfaction can be applied to problems of probabilistic reasoning, social consensus, emotions, and decision making in general. Though his argument about coherence being a theory of everything is not fully convincing, he makes a strong case for specific uses of the theory in concrete problems of decision making and probabilistic reasoning.

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 be fitted together. In this regard, we can see each piece of information as imposing a constraint on another one, the constraints being positive or negative. Positive constraints strengthen pieces of information, thereby increasing coherence, while negative constraints weaken them, thereby increasing incoherence. Hence, we want to put together those pieces of information that have a positive constraint between them, while separating those having a negative constraint. If we manage to partition pieces of information in this manner, then we have satisfied all constraints and we have a state where coherence is maximal. The maximum coherence is achieved when we have satisfied the maximum constraints.

### 3.2.1 Thagard's Formalisation

Thagard formalises coherence as follows: Let $E$ be a finite set of elements $\{e_i\}$ and $C$ be a set of constraints on $E$ understood as a subset $\{(e_i, e_j)\}$ of pairs of elements of $E$. $C$ divides into $C+$, the positive constraints on $E$, and $C-$, the negative constraints on $E$. With each constraint is associated a number $w$, which is the weight (strength) of the constraint. Maximising coherence is formulated as the problem of partitioning $E$ into two sets, $\mathcal{A}$ (accepted) and $\mathcal{R}$ (rejected), in a way that maximises compliance with the following two coherence conditions:

1. if $(e_i, e_j)$ is in $C+$ then $e_i$ is in $\mathcal{A}$ if and only if $e_j$ is in $\mathcal{A}$.
2. if $(e_i, e_j)$ is in $C-$, then $e_i$ is in $\mathcal{A}$ if and only if $e_j$ is in $\mathcal{R}$.

If $(e_i, e_j) \in C$, then, Thagard defines it as a satisfied constraint. If $W$ be the weight of the partition, that is, the sum of the weights of the satisfied constraints. The coherence problem is then to partition $E$ into $\mathcal{A}$ and $\mathcal{R}$ in a way that maximises $W$. Because $a$ coheres with $b$ is a symmetric relation, the order of the elements in the constraints does not matter.

By itself, this characterisation has no philosophical or psychological or probabilistic reasoning applications, because it does not state the nature of the elements, the nature of the constraints, or the algorithms to be used to maximise satisfaction of the constraints.However, Thagard further proposes that there are 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 the algorithms that solve the general coherence problem can be used to compute coherence in ways that apply specific domain problems.

### 3.2.2 Computing Coherence

Since the coherence problem is formalised as a constraint satisfaction problem, he further argues that there should be many algorithms to compute coherence. i.e. we can solve the problem of selecting elements that can be accepted or rejected in a way that maximises compliance with the two coherence conditions on constraint satisfaction. He goes on to give five specific algorithms with increasing degrees of complexity and effectiveness. They are as given below:

1. an exhaustive search algorithm that considers all possible solutions;
2. an incremental algorithm that considers elements in arbitrary order;
3. a connectionist algorithm that uses an artificial neural network to assess coherence;
4. a greedy algorithm that uses locally optimal choices to approximate a globally optimal solution;
5. a semidefinite programming (SDP) algorithm that is guaranteed to satisfy a high proportion of the maximum satisfiable constraints.

Thagard has experimented with many computational implementations of coherence. ECHO is a computational model of explanatory coherence which uses a connectionist algorithm. Though there are 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, which takes us from the abstract notion of coherence to a a computational phenomena which can be evaluated. One of the main drawback of his theory is that, he stops with giving certain principles about calculating values of coherence constraints for different types of coherence. However, to compute these values, one needs to have concrete functions with proven properties. This dissertation is mainly an attempt in this direction, while also attempting to solve problems in normative multiagent systems.

## 3.3 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. That is, everything else is picked and chosen to maximise coherence. In cognitive terms, this would mean that, there are no beliefs nor other cognitions 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. For most researchers of agent theory and multiagent systems, however, changing beliefs in this way is equivalent to creating agents that are not dependable. However we argue that taking decisions based on coherence does not imply an unstable system. Our claim is based on the fact that some beliefs are more fundamental than others. Such fundamental beliefs define a personality, and revision of a fundamental belief is less frequent compared to other beliefs. In coherence terms, these beliefs are fundamental because they support and get support from most other cognitions and hence are in positive coherence with them. Hence, such beliefs will almost always be part of the chosen set while maximising coherence. The same is the case with other cognitions. Those that are normally in positive coherence with most other, will almost always be selected with coherence maximisation, while this process also helps us resolve conflicts by selecting the best alternatives.

When applied to decision making, this means, we not only select the set of actions to be performed to achieve certain fixed goals, but we 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 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 choses goals, actions and even beliefs to fit a grand plan of maximising coherence.

This view of decision making is very different from those of classical decision making theories. The fundamental notion of classical decision theories is the notion

of *preference*. The notion of *utility* is derived from the notion of preference in such a way that $x$ is preferred to $y$ if and only if $x$ has a greater utility than $y$. Then, the decision making process is equated to the maximisation of utilities. However, preferences are atomic, and there is no conceptual understanding of how preferences are formed. In the theory of coherence, we precisely aim at understanding preferences. The assumption here is more basic because the only knowledge available to us are the various interacting constraints between pieces of information. The process of coherence maximisation helps one to form the set of preferences from the available complex network of constraints. Further, coherence-based decision making unlike other multi-attribute decision making processes, works with a dynamic system where everything from beliefs to goals and actions are subject to be selected or discarded. In other theories decision making is more about action selection for a pre-established set of goals based on a pre-established set of criteria.

In this dissertation we discuss how an agent can reason about social norms to aid in decision making, especially when there are conflicts among its cognitions and the norms. However, we attempt to address the more fundamental problem of agent autonomy, and propose a general coherence-based framework that, among other things, helps an agent to reason autonomously about its adherence to social norms, its own goals and beliefs. In this way, we are proposing an extension or an alternative notion of an agent theory where coherence is the fundamental aspect of the agent's cognition which it tries to preserve, and beliefs, goals, intentions and other social dimensions are adjusted to preserve coherence.

# Chapter 4

# A Coherence Framework

In this chapter, we introduce our generic coherence framework together with those computable functions that will allow us to build coherence-based agents (see Section 5). Our framework is based on Thagard's formulation of the theory of coherence as maximising constraint satisfaction. 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 graphs with nodes and edges to model these associations. Here we differ from other approaches in extending agent theories [6, 20] as we rethink the way an agent framework is perceived by making the associations in the cognitions explicit in representation and analysis. That is, we introduce coherence as a fundamental property of the cognition of an agent. In the following definitions, we introduce coherence graphs, the various computable functions to determine the coherence of such graphs, and how the coherence of a given graph can be maximised.

## 4.1 Coherence Graphs

The nodes in a coherence graph represent the pieces of information for which we want to estimate coherence. Examples of such pieces of information are atomic propositions (or complex formulae) and atomic concepts (or their combinations). Nodes have a degree represented by a function $\rho$. This function indicates a confidence associated with the piece of information. For example, if the piece of information is a belief proposition of a BDI agent then the value of $\rho$ for that piece of information indicates the degree to which the agent believes in the proposition.

Finally, edges between nodes may be associated with a strength, represented by a function $\zeta$, which is derived from the underlying relation between the pieces of information. That is, if two pieces of information are related through an *explanation*, for instance, then the function $\zeta$ assigns a positive strength to the edge connecting those pieces of information. Thagard in his characterisation classifies coherence into different types such as *explanatory, deductive, perceptual, conceptual, analogous* and *deliberative coherence* depending on this underlying relation. Thus, we have different

$\zeta$ functions for different types of coherence. The value of the function $\zeta$, that is, the strength on an edge, 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 connecting the pieces of information. Hence we only consider nonzero strength values on edges. A guideline for defining $\zeta$ is Thagard's guiding principles for each type of coherence, which we will elaborate in Section 5 for deductive coherence.

We consider a running example as in Figure 4.1, which will help us to illustrate the concepts as we define them. The graph in the example is constructed with one of the inference rules of the propositional calculus, namely Modus Tollens: $(\alpha \rightarrow \beta), \neg\beta \vdash \neg\alpha$. As we gradually build our framework, we also add more sophistication to our coherence graph in this example.

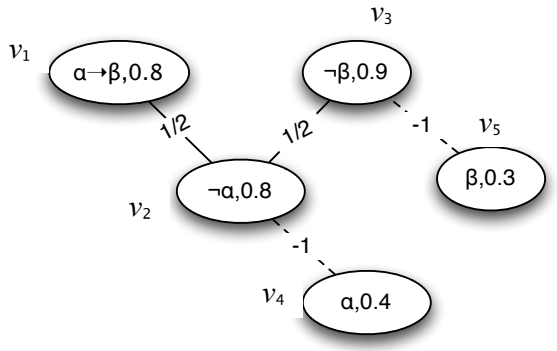


Figure 4.1: Graph representing the coherence and incoherence relations between graded propositions related through Modus Tollens: $(\alpha \rightarrow \beta), \neg\beta \vdash \neg\alpha$

Thus a coherence graph is defined as follows:

**Definition 1** *A* coherence graph *is a graph $g = \langle V, E, \rho, \zeta \rangle$, where*

1. *$V$ is a finite set of nodes representing pieces of information.*
2. *$E$ is a finite set of subsets of 2 elements of $V$ representing the coherence and incoherence between pieces of information.*
3. *$\rho : V \rightarrow [0, 1]$ is a function that maps each node to a weight representing grades (confidence) on the corresponding pieces of information.*
4. *$\zeta : E \rightarrow [-1, 1] \setminus \{0\}$ is a function that assigns a value to the coherence between concepts, and which we shall call a* coherence function[1].

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

Figure 4.1 is an example of a coherence graph as defined above with the following values.

[1]We write indistinguishably $\zeta(v, w)$ or $\zeta(w, v)$ for $\zeta(\{v, w\})$

- $V = \{v_1, v_2, v_3, v_4, v_5\}$
- $E = \{\{v_1, v_2\}, \{v_3, v_2\}, \{v_2, v_4\}, \{v_3, v_5\}\}$
- $\rho(v_1) = 0.8, \rho(v_2) = 0.8, \ldots$
- $\zeta(v_1, v_2) = 0.5, \zeta(v_2, v_4) = -1, \ldots$

## 4.2 Calculating Coherence

According to coherence theory, 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, the 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 2** *Given a coherence graph $g = \langle V, E, \rho, \zeta \rangle$, and a partition $(\mathcal{A}, V \setminus \mathcal{A})$ of $V$, the* set of satisfied constraints $C_{\mathcal{A}} \subseteq E$ *is given by*

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

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

To illustrate this, consider the partition $(\mathcal{A}_1, V \setminus \mathcal{A}_1)$ as in Figure 4.2. We see that, given this partition, the only satisfied constraints are those between $\{v_1, v_2\}$ and between $\{v_2, v_4\}$.

Now we define both the accepted set in the partition that maximises the satisfaction of constraints and the actual value of coherence corresponding to this partition. We define first the *strength of a partition* as the sum over the strengths of all the satisfied constraints ($\zeta$ values) corresponding to that partition, multiplied by the degrees (the $\rho$ values) of the nodes connected by the edge. Then the coherence of a graph is defined to be the maximum among the total strengths when calculated over all its partitions. We have the following definitions:

**Definition 3** *Given a coherence graph $g = \langle V, E, \rho, \zeta \rangle$,* the strength of a partition $(\mathcal{A}, V \setminus \mathcal{A})$ *of $V$ is given by*

$$\sigma(g, \mathcal{A}) = \frac{\sum_{\{v,w\} \in C_{\mathcal{A}}} \mid \zeta(v, w) \mid \cdot \rho(v) \cdot \rho(w)}{\mid E \mid} \tag{4.1}$$

For the partition in Figure 4.2, its *strength* is $0.16$.

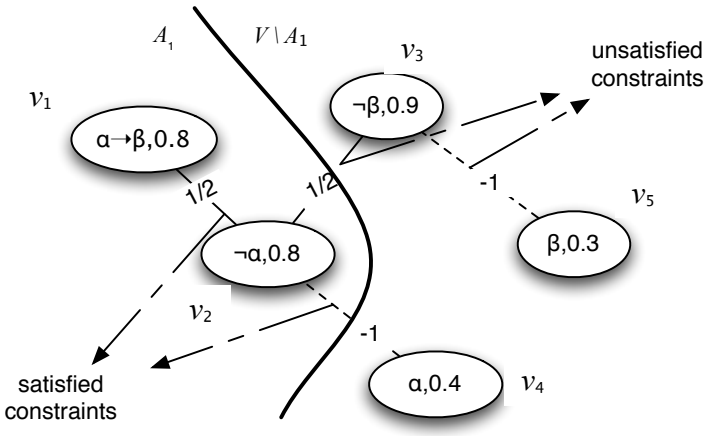


Figure 4.2: The strength of partition $(\mathcal{A}_1, V \setminus \mathcal{A}_1)$ is $0.16$

**Definition 4** *Given a coherence graph $g = \langle V, E, \rho, \zeta \rangle$ and given the strength $\sigma(g, \mathcal{A})$ for all subsets $\mathcal{A}$ of $V$,* the coherence of *$g$ is given by*

$$\kappa(g) = \max_{\mathcal{A} \subseteq V} \sigma(g, \mathcal{A}) \tag{4.2}$$

*If for some partition $(\mathcal{A}, V \setminus \mathcal{A})$ of $V$, the coherence is maximum, that is, $\kappa(g) = \sigma(g, \mathcal{A})$, then the set $\mathcal{A}$ is called the* accepted *set and $V \setminus \mathcal{A}$ the* rejected *set of this partition.*

An important property of coherence maximisation is that, the accepted set $\mathcal{A}$ is not unique. This is due to the fact that the partitions $(\mathcal{A}, V \setminus \mathcal{A})$ and its dual $(V \setminus \mathcal{A}, \mathcal{A})$ are coherence maximising partitions. Hence, whenever $\mathcal{A}$ is a coherence maximising accepted set, so is $V \setminus \mathcal{A}$. Moreover there could be other partitions that generate the same value for $\kappa(g)$. In choosing the preferred accepted set $\mathcal{A}$ from the set of accepted sets, we go back to the theory of coherence and principles set by Thagard.

Thagard's Principle 3 (more discussion on these principles are in Section 5) states that *Propositions that are intuitively obvious have a degree of acceptability on their own.* These obvious propositions are in many cases the evidences or the known facts available to us. Hence, we can choose our accepted set to be the one which includes these obvious propositions. However, one can argue that this does not guarantee uniqueness. Another factor to differentiate the accepted sets is the coherence of the sub-graphs restricted to the accepted sets i.e., $g|_{\mathcal{A}}$. The coherence of the sub-graphs gives us an indication of how strongly connected they are. The higher the coherence, the better connected the pieces of information within the sub-graph. Hence, we should prefer an accepted set corresponding to the sub-graph with a higher coherence to that of a subgraph with a lower coherence. Yet another factor to distinguish the accepted sets are the number of nodes in each set. We should prefer those sets with more nodes as we are interested in eliminating the minimum number of problematic nodes that reduces

our coherence, while trying to retain all that is possible in the accepted set. Apart from these criteria, the solution to preferring one accepted set over another depends on the decision making agent, which can prefer one set to another for independent reasons.

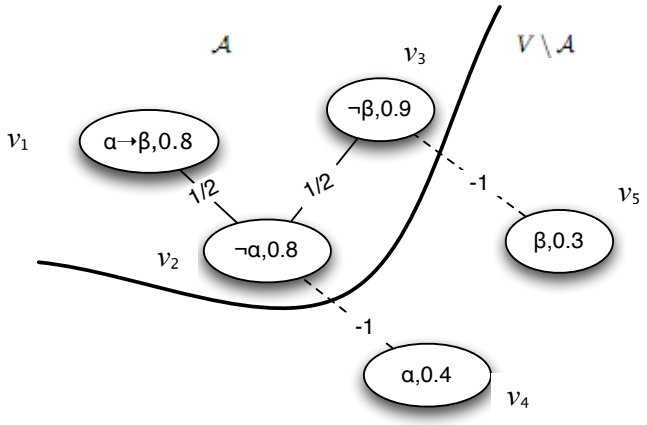


Figure 4.3: Coherence of the graph is $0.3175$ (for the partition $(\mathcal{A}, V \setminus \mathcal{A})$ )

For the example in Figure 4.1, we have a coherence maximising partition $(\mathcal{A}, V \setminus \mathcal{A})$ as in Figure 4.3. With this partition we see that all the constraints are satisfied and this partition gives the maximum strength for the graph.

# Chapter 5

# Formalising Coherence: A Proof-Theoretical Approach

So far we have introduced the general computable functions of our 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. That is, given a set of pieces of information and their associated confidence degrees, we need to define a coherence function $\zeta$ relating them. As the nature of relationship between two pieces of information can vary greatly, we do not have one unique coherence function. Thagard in his characterisation of coherence defines different types of coherence based on the type of pieces of information and their relationships. Further, in each of of these types, only the corresponding relationship is evaluated. That is, in an explanatory coherence, two pieces of information are coherent only if they are related by an explanation. Thagard proposes certain principles to characterise coherence in each of the different types.

Here we study one such coherence, namely deductive coherence, and define a *deductive coherence function* which captures the deductive relationship between propositions. Since logical deduction has a sound theoretical basis, and has well defined rules, we choose deductive coherence among the different types of coherence to start with a formalisation of coherence. We first derive a deductive coherence function in adherence with Thagard's principles and later analyse this function in the context of structural and internal connectives. The latter helps us to further derive coherence values between those pieces of information that are not directly related by deduction.

Thagard introduces in [26] the notion of deductive coherence by means of a set of principles:

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.

In this section we give a proof-theoretical formalisation of the notion of deductive coherence inspired by the principles put forth by Thagard. We base our coherence function on multiset deductive relations. 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 as linear logics, relevance logics or multi-valued logics. We denote a "multiset deductive relation" as MDR. We assume that all MDRs we deal with are finitary and decidable. These MDRs are often called *simple consequence relations* [2]. We define an MDR as follows:

**Definition 5** *Given a logical language $L$, a multiset deductive relation (MDR) on a set of formulas of $L$, is a binary relation $\vdash$ between finite multisets of formulas of $L$ such that, for all $\Gamma_1, \Gamma_2, \Sigma_1, \Sigma_2 \subseteq L$ and for all $\gamma \in L$:*

1. ***Reflexivity**: $\gamma \vdash \gamma$, for every formula $\gamma$*

2. ***Transitivity**: if $\Gamma_1 \vdash \Sigma_1, \gamma$ and $\gamma, \Gamma_2 \vdash \Sigma_2$, then $\Gamma_1, \Gamma_2 \vdash \Sigma_1, \Sigma_2$.*

As usual in sequent calculi, 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.

## 5.1 Coherence Functions

We approach the formalisation of deductive coherence by first deriving a coherence function from an MDR. We use Thagard's principles to relate an MDR and the coherence function $\zeta$. The intuition behind these principles is that whenever two propositions are related by a deductive relation, then there exists a positive coherence between them, the degree of the coherence being inversely proportional to the number of propositions involved in the deduction (Principle 4). If they form a contradiction, then there is a negative coherence between them. However we do not model Principle 3 (propositions that together are used to deduce some other proposition cohere with each other). Principle 3, if taken into account, can mean that all propositions cohere with each other. This is due to certain straightforward deductions such as $\alpha, \beta \vdash \alpha \wedge \beta$. We imagine

that, Thagard, however, had certain specific relationships in mind, such as in cases where premises are implications of one another in the context of the conclusion. These in general are captured by principle 2.

We formalise Thagard's principles in terms of a *support function* $\eta$ on the MDR as below.

**Definition 6** *Let $\vdash$ be an MDR and $\mathcal{T}$ a finite set of formulas of a language $L$. A support function is a partial function $\eta : \mathcal{T} \times \mathcal{T} \to I\!N \cup \{-1\}$ given by*

$$\eta(\alpha,\beta) = \begin{cases} |\Gamma| + 1 & \text{if } \Gamma \subseteq \mathcal{T} \text{ is the smallest set, such that } \Gamma, \alpha \vdash \beta, \text{ and} \\ & \Gamma, \alpha \nvdash \text{ and } \Gamma \nvdash \beta \\ 1 & \text{if } \vdash \beta \text{ and } \alpha \nvdash \\ -1 & \text{if } \alpha, \beta \vdash \\ undefined & \text{otherwise} \end{cases}$$

In our example, we have $(\alpha \to \beta), \neg\beta \vdash \neg\alpha$. That is, there are two premises $\alpha \to \beta$ and $\neg\beta$ required to derive $\neg\alpha$. Therefore, the value of the support function between each of the premises and the conclusion is 2. That is, $\eta(\alpha \to \beta, \neg\alpha) = \eta(\neg\beta, \neg\alpha) = 2$.

Observe that, for any given MDR, the support function $\eta$ satisfies the following:

- If $\alpha \vdash \beta$, then $\eta(\alpha, \beta) = 1$.
- If $\eta(\gamma, \alpha) = 1$ and $\eta(\alpha, \beta) = 1$, then $\eta(\gamma, \beta) = 1$.
- In general, if $\eta(\gamma, \alpha) = n + 1$ and $\eta(\alpha, \beta) = m + 1$ with $n, m \in I\!N$, then $\max(n, m) + 1 \leq \eta(\gamma, \beta) \leq n + m + 1$.

We now define the deductive coherence between two propositions as the value of the stronger relation since deductive coherence is a symmetric function. Due to this, even if there may only be a deductive relation in one direction, there will be a deductive coherence in both directions. The value of deductive coherence is the inverse value given by the support function. 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 7** *Let $\vdash$ be an MDR, $\mathcal{T}$ a finite set of formulas of a language $L$, $\eta : \mathcal{T} \times \mathcal{T} \to I\!N \cup \{-1\}$ be a support function, A deductive coherence function is a partial function $\zeta : \mathcal{T} \times \mathcal{T} \to [-1, 1] \setminus \{0\}$ given by:*

*For any pair $(\alpha, \beta)$ of formulas in $\mathcal{T}$, a* coherence function *$\zeta$ is a partial function with*

$$\zeta(\alpha,\beta) = \begin{cases} 1/\min(\eta(\alpha,\beta), \eta(\beta,\alpha)) & \text{if both } \eta(\alpha,\beta) \text{ and } \eta(\beta,\alpha) \text{ are defined} \\ 1/\eta(\alpha,\beta) & \text{if } \eta(\alpha,\beta) \text{ is defined and } \eta(\beta,\alpha) \text{ undefined} \\ 1/\eta(\beta,\alpha) & \text{if } \eta(\beta,\alpha) \text{ is defined and } \eta(\alpha,\beta) \text{ undefined} \\ undefined & otherwise \end{cases}$$

In our example (Figure 4.1), since we have $\eta(\alpha \to \beta, \neg\alpha) = \eta(\neg\beta, \neg\alpha) = 2$, and since these are the only deduction relations between these formulas, we have $\zeta(\alpha \to \beta, \neg\alpha) = \zeta(\neg\beta, \neg\alpha) = 1/2$. Similarly, since $\eta(\alpha, \neg\alpha) = \eta(\beta, \neg\beta) = -1$, we also have $\zeta(\alpha, \neg\alpha) = \eta(\beta, \neg\beta) = -1$.

**Proposition 5.1.1** *A deductive coherence function $\zeta$ satisfies Thagards's principles except for principle* 3*.*

***Proof***

*Principle*1 : *$\zeta$ is symmetric by construction.*

*Principle* 2 : *If $|\Gamma| = n$ and $\Gamma, \alpha \vdash \beta$ then $\eta(\alpha, \beta) = n + 1$ and $\zeta(\alpha, \beta) = 1/(n + 1)$.*

*Principle* 4 : *If $|\Gamma_1| = n$ and $|\Gamma_2| = m$, $n < m$ and $\Gamma_1, \alpha_1 \vdash \beta$ and $\Gamma_2, \alpha_2 \vdash \beta$ then $\zeta(\alpha_1, \beta) = 1/(n + 1) > \zeta(\alpha_2, \beta) = 1/(m + 1)$.*

*Principle* 5 : *Satisfied by construction.*

*Principle* 6 : *Propositions that are intuitively obvious are the axioms. That is, if $\vdash \beta$ is an axiom, then for every $\alpha \nvdash$, we have $\zeta(\alpha, \beta) = 1$. That is, $\beta$ coheres with every other proposition with the highest coherence. Hence $\beta$ has an intuitive priority.*

*Principle* 7 : *Satisfied by the definition 4.2.*

Note that deductive relations are not symmetric but are transitive in general. However coherence functions differ by not being transitive in general. This is due to the symmetric property of a coherence function. That is, a deductive relation in a single direction gives raise to a coherence function in both directions. However, if we exclude certain special cases, we can show that coherence functions are transitive.

**Proposition 5.1.2** *Whenever $\alpha, \beta, \gamma \nvdash$ and $\nvdash \alpha$, $\nvdash \beta$, $\nvdash \gamma$,*

$$C(\gamma, \alpha) = 1 \text{ and } C(\alpha, \beta) = 1, \text{ then } C(\gamma, \beta) = 1$$

*except for the two following cases for non-equivalent formulas:*

- $\gamma \vdash \alpha$ *and* $\beta \vdash \alpha$
- $\alpha \vdash \gamma$ *and* $\alpha \vdash \beta$

## 5.2 Properties of Coherence Based On MDRs

We can classify logics according to structural rules (such as weakening /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 the internal and combining connectives, and makes no difference between the combining and the corresponding internal connectives. On the other hand, propositional linear logic is nonmonotonic, has the above connectives but distinguishes between internal and combining ones. Intuitionistic logic differs from classical propositional logic in its implication connective and does not contain any internal negation. In this section, we explore the properties of the deductive coherence function $\zeta$ which would help determine the value of function $\zeta$ between pairs of formulas which are related through some

of the structural rules and connectives. We do this by identifying the properties of the support function $\eta$ using the properties of the connectives and structural rules.

By Definition 6, the function $\eta$ is defined for formulas 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 $\eta$ between different formulas of the premises and conclusion, using the properties of the connectives.

### 5.2.1 Combining Conjunction

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

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

Consequently, for all $\Gamma \subseteq L$ and $\alpha, \beta, \gamma \in L$ and $n, m \geq 0$

1. Given that $\Gamma, \gamma \vdash \alpha \wedge \beta$ implies $\Gamma, \gamma \vdash \alpha$ and $\Gamma, \gamma \vdash \beta$, we have that
   if $\eta(\gamma, \alpha \wedge \beta) = n + 1$ then $0 < \eta(\gamma, \alpha) \leq n + 1$ and $0 < \eta(\gamma, \beta) \leq n + 1$.
2. Given $\Gamma, \gamma \vdash \alpha$ and $\Gamma, \gamma \vdash \beta$ implies $\Gamma, \gamma \vdash \alpha \wedge \beta$, and $\vdash$ satisfies weakening, then we have that
   if $\eta(\gamma, \alpha) = n + 1$ and $\eta(\gamma, \beta) = m + 1$ then $\max(n, m) + 1 \leq \eta(\gamma, \alpha \wedge \beta) \leq n + m + 1$
3. Given $\alpha, \beta \nvdash$ then we have that
   $\eta(\alpha \wedge \beta, \alpha) = 1$ and $\eta(\alpha \wedge \beta, \beta) = 1$

### 5.2.2 Internal Conjunction

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

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

Consequently, for all $\Gamma \subseteq L$ and $\alpha, \beta, \sigma \in L$ and $n, m \geq 0$

1. Given that $\Gamma, \alpha \circ \beta \vdash \sigma$ implies $\Gamma, \alpha, \beta \vdash \sigma$
   if $\eta(\alpha \circ \beta, \sigma) = n + 1$ implies $0 < \eta(\alpha, \sigma) \leq n + 2$ and $0 < \eta(\beta, \sigma) \leq n + 2$
2. Given that $\Gamma, \alpha, \beta \vdash \sigma$ implies $\Gamma, \alpha \circ \beta \vdash \sigma$ and that $\vdash$ satisfies weakening, we have that
   if $\eta(\alpha, \sigma) = n+1$ and $\eta(\beta, \sigma) = m+1$ implies $max(n, m)+1 \leq \eta(\alpha \circ \beta, \gamma) \leq n + m + 1$
3. Given that $\alpha, \beta \vdash$ iff $\alpha \circ \beta \vdash$, then we have that
   if $\eta(\alpha, \beta) = -1$ then, for all $\gamma \in L$ we have $\eta(\gamma, \alpha \circ \beta) = -1$
4. Given $\alpha, \beta \nvdash$ then we have that
   $\eta(\alpha, \alpha \circ \beta) = 2$ and $\eta(\beta, \alpha \circ \beta) = 2$

### 5.2.3 Combining Disjunction

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

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

Consequently, for all $\Gamma \subseteq L$ and $\alpha, \beta, \sigma \in L$ and $n, m \geq 0$

1. Given that $\Gamma, \alpha \vee \beta \vdash \sigma$ implies $\Gamma, \alpha \vdash \sigma$ and $\Gamma, \beta \vdash \sigma$, we have that
   if $\eta(\alpha \vee \beta, \sigma) = n + 1$ then $0 < \eta(\alpha, \sigma) \leq n + 1$ and $0 < \eta(\beta, \sigma) \leq n + 1$
2. Given that $\Gamma, \alpha \vdash \sigma$ and $\Gamma, \beta \vdash \sigma$ implies $\Gamma, \alpha \vee \beta \vdash \sigma$ and that $\vdash$ satisfies weakening, we have that
   if $\eta(\alpha, \sigma) = n + 1$ and $\eta(\beta, \sigma) = m + 1$ then $max(n, m) + 1 \leq \eta(\alpha \vee \beta, \sigma) \leq n + m + 1$
3. Given that $\gamma, \alpha \vee \beta \vdash$ iff $\gamma, \alpha \vdash$ and $\gamma, \beta \vdash$, then we have that
   if $\eta(\gamma, \alpha \vee \beta) = -1$ iff $\eta(\gamma, \alpha) = -1$ and $\eta(\gamma, \beta) = -1$
4. Given $\alpha, \beta \nvdash$ then we have that
   $\eta(\alpha, \alpha \vee \beta) = 1$ and $\eta(\beta, \alpha \circ \beta) = 1$

### 5.2.4 Internal Disjunction

Disjunction $+$ is *internal* iff, for all $\Gamma, \Sigma \subseteq L$ and $\alpha, \beta \in L$ we have:

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

Consequently, for all $\Gamma \subseteq L$ and $\alpha, \beta, \gamma \in L$ and $n, m \geq 0$

1. Given that $\Gamma, \gamma \vdash \alpha + \beta$ implies $\Gamma, \gamma \vdash \alpha, \beta$ and that $\vdash$ satisfies weakening, we have that
   if $\eta(\gamma, \alpha) = n + 1$ and $\eta(\gamma, \beta) = m + 1$ then $\eta(\gamma, \alpha + \beta) = min(n, m) + 1$
   if $\eta(\gamma, \alpha) = n + 1$ and if $\eta(\gamma, \beta)$ undefined, then $\eta(\gamma, \alpha + \beta) = n + 1$
2. Given $\alpha, \beta \nvdash$ and that $\vdash$ satisfies weakening, then we have that
   $\eta(\alpha, \alpha + \beta) = 1$ and $\eta(\beta, \alpha + \beta) = 1$

### 5.2.5 Combining Implication

Implication $\supset$ is *combining* iff for all $\Gamma, \Sigma \subseteq L$ and $\alpha, \beta \in L$ we have:

$$\Gamma, \alpha \supset \beta \vdash \Sigma \text{ iff } \Gamma \vdash \Sigma, \alpha \text{ and } \Gamma, \beta \vdash \Sigma.$$

Consequently, for all $\Gamma \in L$ and $\alpha, \beta, \sigma \in L$ and $n, m \geq 0$

1. Given that $\Gamma, \alpha \supset \beta \vdash \sigma$ implies $\Gamma, \beta \vdash \sigma$ we have that
   if $\eta(\alpha \supset \beta, \sigma) = n + 1$ then $\eta(\beta, \sigma) = n + 1$

2. Given that $\gamma, \alpha \supset \beta \vdash$ iff $\gamma \vdash \alpha$ and $\gamma, \beta \vdash$, we have that
   if $\eta(\gamma, \alpha \supset \beta) = -1$ iff $\eta(\gamma, \alpha) = 1$ and $\eta(\gamma, \beta) = -1$

3. Given $\alpha, \beta \nvdash$ then we have that
   $\eta(\beta, \alpha \supset \beta) = 1$
   and if $\vdash$ satisfies weakening, then we have that
   $\eta(\alpha, \beta) = 1$ and $\eta(\alpha \supset \beta, \beta) = 1$

### 5.2.6 Internal Implication

Implication $\rightarrow$ is *internal* iff for all $\Gamma, \Sigma \subseteq L$ and $\alpha, \beta \in L$ we have

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

Consequently, for all $\Gamma \in L$ and $\alpha, \beta, \gamma \in L$ and $n, m \geq 0$

1. Given that $\Gamma, \gamma, \alpha \vdash \beta$ iff $\Gamma, \gamma \vdash \alpha \rightarrow \beta$, we have that
   if $\eta(\gamma, \alpha \rightarrow \beta) = n + 1$ then $0 < \eta(\gamma, \beta) \leq n + 2$
   if $\eta(\gamma, \beta) = n + 2$ then $0 < \eta(\gamma, \alpha \rightarrow \beta) \leq n + 1$

2. Given $\alpha, \beta \nvdash$ we have that
   $\eta(\alpha \rightarrow \beta, \beta) = 2$

### 5.2.7 Internal Negation

Negation is *internal* iff, for all $\Gamma, \Sigma \subseteq L, \alpha \in L$ we have:

$$\Gamma, \alpha \vdash \Sigma \text{ iff } \Gamma \vdash \Sigma, \neg\alpha$$

Consequently, for all $\Gamma \subseteq L$ and $\alpha, \gamma \in L$

1. if $\alpha \nvdash$ and $\nvdash \alpha$ and given that $\gamma, \alpha \vdash$ iff $\gamma \vdash \neg\alpha$, we have that,
   $\eta(\gamma, \alpha) = -1$ iff $\eta(\gamma, \neg\alpha) = 1$

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

An interesting point to note is that, most properties we have listed in this section hold universally, however, a few properties need that the deduction relation satisfies weakening. This has a special significance for deductive coherence as the principles of deductive coherence indirectly assumes the absence of weakening. That is, two propositions are related by deductive coherence only if one of them contributes in deriving the other. When weakening is introduced, this constraint no longer holds. Hence coherence is more closer in structure to non-classical logics such as relevant logic where the antecedent needs to be necessarily relevant to the consequent.

## 5.3 An Example

So far we have analysed formally the properties of coherence by listing the properties of the support function in terms of the connectives and structural properties of a logic. Here we apply these properties in the context of classical logic to deduce some of the coherence values. We use the same example as in the previous sections. We enrich the example by adding another proposition $\gamma$. This is because of the fact that, the more interesting properties of the support function are derived between distinguished elements and non-distinguished elements. We also add an implication namely $\gamma \vdash \alpha \rightarrow \beta$. Then using the axioms we have derived so far, we can deduce the following coherence values. Since in the example we only have implications and negations, we use only the axioms related to implication and negation. It is however easy to see how we can similarly apply the results of other connectives in appropriate cases. The purpose of this example is only to demonstrate that coherence graphs can be enriched with these properties of coherence, however, we do not intend to be exhaustive.

1. $\zeta(\beta, \alpha \rightarrow \beta) = 1/2$ using Property 2 (Internal Implication).
2. $\zeta(\beta, \gamma) = 1/2$ using Property 1 and Property 2 (Internal Implication).

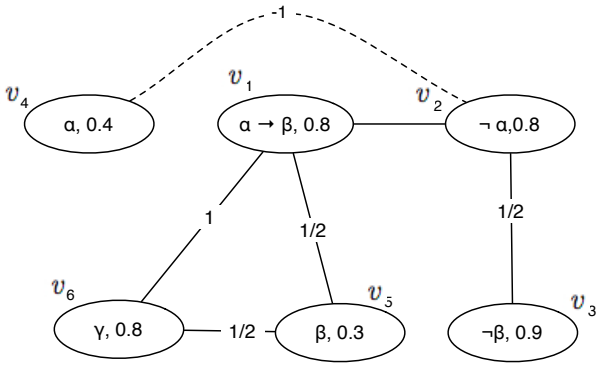


Figure 5.1: Applying properties of $\eta$ to compute coherence values

# Chapter 6

# An Architecture for Coherence-Driven Agents

In this chapter we describe an architecture for coherence-driven agents based on the coherence framework developed so far. A *coherence-driven agent* is an agent which always takes an action based on maximisation of its coherence. We further consider cognitive agents such as those based on the BDI theory, since it is one of the prominent existing agent architectures. We use an adaptation of the architecture developed by Casali et al. [7] based on multi-context systems (MCS), which incorporates graded cognitions. The grade in a cognition represents the degree to which an agent believes (desires or intents) a particular cognition. We use graded cognitions to incorporate reasoning under uncertainty into our agent framework. Then, an MCS models the representation and interaction between these graded cognitions.

In the work of Casali et al., the MCS specification of an agent contains three basic components: units or contexts, logics, and bridge rules, which channel the propagation of consequences among the contexts. Contexts in a multi-context BDI are the *mental* contexts of beliefs, desires, and intentions. The deduction mechanism of MCS is based on two kinds of inference rules, internal rules $\Delta_i$ inside each context, and bridge rules $B$ 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 [14, 15]. Thus, an agent is defined as a family of interconnected contexts:

$$\left\langle \{C_i\}_{i \in I}, B \right\rangle$$

where

- each context $C_i = \langle L_i, A_i, \Delta_i \rangle$ consists of
  - $L_i$ are languages
  - $A_i$ are axioms
  - $\Delta_i$ are deduction rules

- $B$ of inference rules with premises and conclusions in different contexts.

  For instance:

$$\frac{1:\psi, 2:\varphi}{3:\phi}$$

  represents that if formula $\psi$ is deduced in context $C_1$ and formula $\varphi$ is deduced in context $C_2$ then formula $\phi$ is inferred in context $C_3$.

The multi-context architecture is adapted here so that we have further structure in the form of coherence graphs, associated with each of our contexts. Our bridge rules carry deductions from one graph to another. As we need the specific coherence graphs to define our multi-context architecture, we first define those specific coherence graphs which are of interest to us. We also need an adaptation of the bridge rules to carry deductions across graphs. After defining these necessary elements, we discuss the agent architecture itself.

## 6.1 Cognitive and Norm Coherence graphs

Given the general definition of a *coherence graph* and its restriction into a *deductive coherence graph* in Chapter 4 and Chapter 5, we here discuss certain specific coherence graphs, the *belief, desire, intention, and norm* coherence graphs of an agent. The graphs will be deductive coherence graphs, since so far we have only defined a deductive coherence function. Further, the nodes of the graph will be elements of logics corresponding to cognitions. Hence, we first define the underlying logic, and later define the graph over this logic. These logics are defined as in Casali et.al. We briefly describe the belief logic below, for a detailed discussion on desire and intention logics, we refer to [7].

In order to define a *belief graph*, we need to first define a belief logic $\mathcal{K}_B = \langle L_B, A_B, \vdash_B \rangle$. We define the belief language $L_B$ by extending the classical propositional language $L$ defined upon a countable set of propositional variables $PV$ and connectives $(\neg, \rightarrow)$. We extend $L$ with a fuzzy unary modal operator B. The modal language $L_B$ is built from the elementary modal formulae $B\varphi$ where $\varphi$ is propositional, and truth constants $r$, for each rational $r \in Q \cap [0,1]$, 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 $\mathcal{K}_B$[1]. Formally the belief language $L_B$ is defined as:

**Definition 8** *Given a propositional language L, a belief language $L_B$ is given by:*

- *if $\varphi \in L$, and $r \in Q \cap [0,1]$ then $(B\varphi, r) \in L_B$*
- *if $\varphi, \psi \in L_B$ then $\varphi \rightarrow_L \psi \in L_B$ and $\varphi \,\&\, \psi \in L_B$ (where $\&$ and $\rightarrow_L$ correspond to the conjunction and implication of Łukasiewicz logic)*
- *if $\varphi \in L$ then $\neg_L \varphi \in L_B$*

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

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

1. All axioms of propositional logic.

2. Axioms of Łukasiewicz logic for modal formulas (for instance, axioms of Hájek's Basic Logic (BL) [16] plus the axiom: $\neg\neg\Phi \rightarrow \Phi$.)

3. Probabilistic axioms, given $\varphi, \psi \in L$ :

   - $B(\varphi \rightarrow \psi) \rightarrow_L (B\varphi \rightarrow B\psi)$
   - $B\varphi \equiv \neg_L B(\varphi \wedge \neg\psi) \rightarrow_L B(\varphi \wedge \psi)$
   - $\neg_L B\varphi \equiv B\neg\varphi$.

The deduction rules defining $\vdash_B$ of $\mathcal{K}_B$ are:

1. Modus ponens.

2. Necessitation for $B$ (from $\varphi$ derive $B\varphi$).

A belief graph over the belief logic $\mathcal{K}_B$ is then defined as follows:

**Definition 9** *Given a belief logic $\mathcal{K}_B = \langle L_B, A_B, \vdash_B \rangle$ where $L_B$ is a belief language, $A_B$ are a set of axioms and $\vdash_B$ are a set of deduction rules, a belief graph $g_B = \langle V_B, E_B, \rho_B, \zeta_B \rangle$ is a coherence graph defined over $\vdash_B$ and a finite set $\mathcal{T}_B \subseteq L_B$ of formulas such that:*

- *$V_B \subseteq \mathcal{T}_B$*
- *$E$ is a set of subsets of 2 elements of $V_B$*
- *$\rho_B : V_B \rightarrow [0,1]$ is defined by means of the truth-functions of Łukasiewicz logic and the probabilistic interpretation of beliefs as follows:*
  - *$\rho(B\varphi, r) = r$ for all $r \in Q \cap [0,1]$*
  - *$\rho(\varphi \,\&\, \psi) = max(\rho(\varphi) + \rho(\psi) - 1, 0)$ for all $\varphi, \psi \in L_B$*
  - *$\rho(\varphi \rightarrow_L \psi) = min(1 - \rho(\varphi) + \rho(\psi), 1)$ for all $\varphi, \psi \in L_B$*
- *$\zeta_B$ is the deductive coherence function defined over $\vdash_B$ and $\mathcal{T}_B$.*

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

A belief graph exclusively represents the graded beliefs of an agent and the associations among them. A desire graph ($g_D$), and an intention graph ($g_I$) over given logics $L_D$, and $L_I$ respectively would be similarly defined. (Analogously the set of all desire, and intention graphs are $\mathcal{G}_D$, and $\mathcal{G}_I$ respectively.)

### 6.1.1 Norm Graph ($g_N$)

The normative behaviour in a normative multiagent system is generally described by using deontic constraints, such as obligations, permissions and prohibitions. Just as we have graded cognitions for an agent, our norms also come with grades. Grades in general add more richness to the semantics, and, in particular for the case of norms, the grades help 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 the social goals it helps in achieving. However, there could be other measures for determining priority of a norm.

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 we have graded norms, we define $\mathcal{K}_N$ as a graded deontic logic namely the Probability-valued Deontic Logic [10] to represent and reason with norms. We define the norm language $L_N$ by extending the classical propositional language $L$ defined upon a countable set of propositional variables and connectives $(\neg, \rightarrow)$. $L_N$ is defined as a fuzzy modal language over Standard Deontic Logic (SDL) to reason about the probability degree of deontic propositions. In our case the probability values are replaced by the grades associated with the norms. The language, axioms and deductions rules are defined similarly as in the case of the belief logic. For the details, refer [10].

Here we list some of the examples of valid norms in PSDL. Below we use triples form for propositional formulas. (The triples in the examples are self explanatory).

- $(\langle John, use, publictransport \rangle \rightarrow O\langle John, validate, ticket \rangle, 0.8)$

  If John uses public transport, then John is obliged to validate the ticket.

- $(\langle Anna, citizen, Utopia \rangle \rightarrow O\langle Anna, pay\_tax, Utopia \rangle, 1)$

  If Anna is a citizen of Utopia, then Anna ought to pay tax to Utopia.

## 6.2 Graph Composition

From our construction so far, a coherence-driven agent will only have separate graphs corresponding to each of its cognitions and to the norm of the different normative systems. Though it is useful to reason within each of the individual cognitions and normative systems, what is more interesting is to be able to reason with all cognitions and norms put together. For instance, it is desirable to know the most coherent set of beliefs, desires, intentions and norms, or to choose the best set of norms and intentions given a set of beliefs and desires. In summary, it will be useful for an agent to reason by taking into account the influence of all its cognitions and external entities such as norms. If an agent is capable of determining the coherence graph of all its cognitions and norms, then the coherence maximising partition for that graph would for instance, tell the agent, which of its intentions are most coherent with a given set of beliefs and desires. Furthermore, to an agent that is part of a normative system, it would tell the agent whether certain norms are coherent with its cognitions. This might help an agent to know whether to obey or violate a norm. this leads naturally to the notion of

reasoning across graphs, and to explore the coherence or incoherence that might exist between nodes of different graphs.

In a multi-context system, reasoning across context is done by the help of bridge rules. Here our contexts are not merely sets of cognitions, but coherence graphs. That is, we need an inference mechanism that relates individual coherence graphs. We achieve this with the help of certain composition functions that connect individual graphs by taking into account the influence of graphs on each other. Later in the section, we take bridge rules as an example to show how we can reason across graphs using composition functions applied over bridge rules.

### 6.2.1 Composition Functions

We break down the graph composition in two phases. The first deals with extending some of the graphs by due to the influence of other graphs in the composition. The second phase deals with joining these graphs by linking nodes of different graphs. For the purpose, we define two kinds of functions, graph node extension functions (denoted with $\varepsilon$) and edge extension functions (denoted with $\iota$) to achieve the composition of graphs.

The first phase of the graph composition takes into account the influence of graphs on each other. Let's assume, for instance, that an agent wants it to be the case that whenever there the agent has an intention $(I\varphi, r)$ in its intention graph, then it wants to infer the corresponding belief $(B\varphi, r)$ in the belief graph. The extension function models such scenarios.

**Definition 10** *Given $n > 0$, we say that a function $\varepsilon : \mathcal{G}^n \to \mathcal{G}^n$ is a* graph extension function *if, given a tuple of graphs $\bar{g} = \langle g_1, \cdots, g_n \rangle$ in $\mathcal{G}^n$, $\varepsilon(\bar{g}) = \bar{g}'$ is such that*

- $V_i' \supseteq V_i$
- $E_i' = E_i$
- $\rho_i'|_{V_i} : V_i \to [0, 1]$ *such that* $\rho_i'|_{V_i}(B\varphi, r) = r$ *where* $(B\varphi, r) \in V_i$
- $\zeta_i' = \zeta_i$.

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

One of the desirable properties of $\mathcal{E}$ is the existence of a *fixed point*. This is because the fixed point would give us a terminating condition for the repeated application of extension functions. We call $\bar{h}$ a fixed point of a subset of $\mathcal{E}$ if the repeated application of its extension functions does not change $\bar{h}$.

**Definition 11** *Given $n, j > 0$, we say that a sequence is an* extension sequence *if, given a tuple of graphs $\bar{g} \in \mathcal{G}^n$ and a set of functions $S \subseteq \mathcal{E}$,*

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

*and say that the elements of $g^j$ 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 $|S^*(\bar{g})| = 1$.*

The second phase of the composition joins all graphs together and adds additional edges between the nodes of the original graphs. Consider our previous example and let's assume that our agent further wants it the case that, the belief and the intention nodes are related and have a positive coherence between them. The function $\iota$ takes $n$ graphs and joins the graphs by adding new edges between related nodes. We have the following definition for $\iota$:

**Definition 12** *Given $n > 0$, we say that a function $\iota : \mathcal{G}^n \to \mathcal{G}$ is a* graph join function *if given a tuple of graphs $\bar{g}$ in $\mathcal{G}^n$, $\iota(\bar{g}) = \langle V, E, \rho, \zeta \rangle$ such that:*

- $V = \bigcup_{1 \leq i \leq n} \{\langle \varphi, i \rangle \mid \varphi \in V_i\}$
- $E \supseteq \bigcup_{1 \leq i \leq n} \{\{\langle \varphi, i \rangle, \langle \psi, i \rangle\} \mid \{\varphi, \psi\} \in E_i\}$
- $\rho : V \to [0, 1]$ such that $\rho(\langle \varphi, i \rangle) = \rho_i(\varphi)$
- $\zeta : E \to [-1, 1]$ such that $\zeta(\langle \varphi, i \rangle, \langle \gamma, i \rangle) = \zeta_i(\varphi, \gamma)$

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

Now we define the composition of graphs in a tuple $\bar{g}$ by combining the two functions $\varepsilon$ and $\iota$. That is, we apply the set of functions $T \subseteq \mathcal{J}$ on the fixed point of the set of functions $S \subseteq \mathcal{E}$ applied over $\bar{g}$. The union of all the resulting graphs is defined as a composition of graphs. Note that here we assume $S$ has a unique fixed point applied over any tuple of graphs $\bar{g}$. It is however a fare assumption given that we can construct the functions in $S$ and $T$ according to the requirements.

**Definition 13** *Given $n > 0$, we say that a function $\varsigma : \mathcal{G}^n \to \mathcal{G}$ is a* graph composition function *if, given a tuple of graphs $\bar{g}$ in $\mathcal{G}^n$, a set of functions $S \subseteq \mathcal{E}$ with a unique fixed point and a set of functions $T \subseteq \mathcal{J}$ then $\varsigma(\bar{g}) = \bigcup_{\iota \in T} \iota(S^*(\bar{g}))$.*

### 6.2.2 Bridge Rules — A Set of Composition Functions

Now we define one such set of graph composition functions by means of a set of bridge rules. Bridge rules have been traditionally used to make inferences across contexts. Here we extend the use of it to make coherence associations across graphs.

**Definition 14** *Given $n > 0$, a bridge rule $b$ is a rule of the form*

$$\frac{i_1 : (A_1, r_1), i_2 : (A_2, r_2), \cdots, i_q : (A_q, r_q)}{j : (A, f(r_1, r_2, \cdots, r_q))}$$

*with:*

- $1 \leq i_k \leq n$, $1 \leq k \leq q$
- $1 \leq q \leq n$

- $1 \leq j \leq n$
- *$A$, $A_k$ are formula schemata and $r$, $r_k \in [0,1]$*
- *$f : [0,1]^q \rightarrow [0,1]$ where $1 \leq q \leq n$*

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

Given a bridge rule $b \in \mathcal{B}$, we derive a pair of functions from them. The first function is from the set $\mathcal{E}$ and we define it as extending the graph in the position $j$ in the tuple with a new formula node represented by the formula schemata $A$. The second function is from the set $\mathcal{J}$ and we define it as a set of coherence functions between formulas represented by the schemata $A_k$ and $A$. That is, we extend the coherence function $\zeta$ to make inferences across graphs.

**Definition 15** *Given a tuple of graphs $\bar{g} = \langle g_1, g_2, \cdots, g_n \rangle$, a bridge rule*

$$\frac{i_1 : (A_1, r_1), i_2 : (A_2, r_2), \cdots, i_q : (A_q, r_q)}{j : (A, f(r_1, r_2, \cdots, r_q))}$$

*as in Definition 14 and if, for all $k$ we have $(\pi(A_k), r_k) \in V_k$, where $\pi$ is the most general substitution making the formula schemata $A_k$ match nodes in $V_k$, then an extension function $\varepsilon$ and a join function $\iota$ are derived from $b$ as:*

1. *$\varepsilon(\bar{g}) = \langle g_1, g_2, \cdots, g'_j, \cdots, g_n \rangle$ where $g'_j = \langle V'_j, E'_j, \rho'_j, \zeta'_j \rangle$ with*
   - $V'_j = V_j \cup \{\pi(A), f(r_1, r_2, \cdots, r_q)\}$
   - $E'_j = E_j$
   - $\rho'_j(v) = \rho_j(v)$ *for all* $v \in V_j$
     $\rho'_j((\pi(A_k), f(r_1, r_2, \cdots, r_q))) = f(r_1, r_2, \cdots, r_q)$ *for all* $1 \leq k \leq q$
   - $\zeta'_j(v, w) = \zeta_j(v, w)$ *for all* $v, w \in V_j$

2. *$\iota(\bar{g}) = \langle V, E, \rho, \zeta \rangle$ as in Definition 12 such that:*
   *For all* $1 \leq k \leq q$, *and* $1 \leq m \leq q$*:*
   - $V = \bigcup_{i \neq j} V_i$
   - $\{\pi(A_k), \pi(A)\} \in E$;
   - $\{\pi(A_k), \pi(A_m)\} \in E$;
   - $\eta(\pi(A_k), \pi(A)) = q + 2$;
   - $\eta(\pi(A_k), \pi(A_m)) = q + 2$
   - $\zeta(v, w) = 1/min\{\eta(v, w), \eta(w, v)\}$ *for all* $v, w \in V$

### 6.2.3 Application of Composition Functions — An Example

Here we consider one such bridge rule and derive both a graph extension function $\varepsilon$ and a graph join function $\iota$ in the context of generating a composition of graphs $g_B \in \mathcal{G}_B$, $g_D \in \mathcal{G}_D$, and $g_I \in \mathcal{G}_I$.

1. Given a bridge rule
$$b = \frac{1 : (B\psi, r), 2 : (D\psi, s)}{3 : (I\psi, \min(r, s))}$$
Where the indices $1, 2$ and $3$ correspond to the contexts $C_1, C_2$, and $C_3$ which are associated with the graphs $g_B, g_D$ and $g_I$ respectively.

2. And Given $(B\psi, 0.95) \in g_B$, $(D\psi, 0.95) \in g_D$

Applying the graph extension function $\varepsilon : \mathcal{G}^n \to \mathcal{G}^n$, we update the graph $g_I$ as below.

- $V_I = V_I \cup (I\psi, 0.95)$
- $\rho_I(I\psi, 0.95) = 0.95$

Now applying the graph join function $\iota : \mathcal{G}^n \to \mathcal{G}$, we update the composition graph $g = \langle V, E, \rho, \zeta \rangle$ which is the composition of the graphs $g_B, g_D$ and $g_I$ as below.

- $E = E_B \cup E_D \cup E_I \cup \{\{(I\psi, 0.95), (B\psi, 0.95)\}, \{(I\psi, 0.95), (D\psi, 0.95)\}\}$
- $\zeta((I\psi, 0.95), (B\psi, 0.95)) = 0.33$;
  $\zeta((I\psi, 0.95), (D\psi, 0.95)) = 0.33$

## 6.3 Coherence-driven Agents

Given the cognitive and norm coherence graphs and the composition functions, we can now turn our attention to 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 $C_i = \langle L_i, A_i, \Delta_i \rangle$ where $L_i$, $A_i$ and $\Delta_i$ are the language, axioms, and inference rules respectively. In our extension of MCS, a coherence-driven agent will further have a function $f$ that maps the set of contexts to a set of corresponding coherence graphs. And a function $h$ that maps a set of bridge rules to a set of graph composition functions. These extensions are due to the introduction of coherence graphs into the contexts. An agent will need both the graph construction functions and a set of deduction mechanisms to reason within and between graphs. For the BDI agents considered here, the contexts are $C_1$, $C_2$, $C_3$ and $C_4$ which determine a belief graph $g_B$, a desire graph $g_D$, an intention graph $g_I$ and a norm graph $g_N$ respectively. Hence we have the following definition:

**Definition 16** *A* coherence-driven agent *a is a tuple $\langle \{C_i\}_{1 \leq i \leq 4}, B, f, h \rangle$ where $\{C_i\}$ is a family of contexts, $B \subseteq \mathcal{B}$ is a set of bridge rules, $f : \{C_i\} \to \mathcal{G}$ maps contexts to coherence graphs, and $h : B \to \mathcal{E} \times \mathcal{J}$ maps bridge rules to pairs of graph extension and graph join functions.*

In the following we describe the process that a coherence-driven agent follows in a normative multi agent system. A coherence-driven agent always tries to maximise coherence, and takes each action based on how it helps in maximising coherence. However, here we limit our discussions to analysing the agent's reasoning with respect to only those actions that are normative, since we are interested in the normative behavior of the agent. We consider one basic normative action of the agent, namely *norm evaluation*. Norm evaluation corresponds to assessing the norm with regard to the situation of the agent, its beliefs, and the goals it wants to achieve. The two situations in which we consider norm evaluation as necessary are *norm adoption* and *norm compliance or violation*. The need for norm adoption occurs when:

1. an agent joins a normative system;
2. a normative system it is part of, tries to modify an existing norm;
3. a normative system it is part of, proposes a new norm for adoption.

Norm compliance or violation occurs when:

1. there is a change in the external environment (such as the joining of new agents in the normative system);
2. there is a change in the internal cognition of the agent (such as adopting a new intention).

Ignoring the conceptual difference in evaluations in both the above cases, we concentrate on how norm evaluation is carried out. A coherence-driven agent evaluates a norm by evaluating the coherence of the composite graph of its cognitions and norms. Below we detail the norm evaluation process when, in particular, the agent encounters a new graded belief $(B\varphi, r)$ (either communicated to the agent by others, by observation, or internally deduced). Other situations discussed above can be similarly evaluated.
Below we describe the procedure a coherence-driven agent follows when it encounters a new belief. We assume the following input for the procedure. To simplify the algorithm, we assume just one bridge rule as part of the input.

- norm to be evaluated $n$;
- the new belief $(B\varphi, r)$;
- the cognitive and norm coherence graphs of the agent $g_B$, $g_D$, $g_I$, and $g_N$;
- a bridge rule $b \in \mathcal{B}$.

1: $v := (B\varphi, r)$
2: $V_B := V_B \cup \{v\}$
3: $\rho_B(v) := r$
4: **for** all $w \in V_B$ **do**
5: compute $\zeta(v, w)$ using Definition 7.
6: **if** $\zeta(v, w)$ is defined **then**

7: $E_B := E_B \cup \{\{v, w\}\}$
8: **end if**
9: **end for**
10: $\bar{g} := \langle g_B, g_D, g_I, g_N \rangle$
11: $\bar{g}' := \varepsilon(\bar{g})$
12: **while** $\bar{g} \neq \bar{g}'$ **do**
13: $\bar{g} := \bar{g}'$
14: $\bar{g}' := \varepsilon(\bar{g})$ where $\varepsilon$ as in Definition 10 derived from bridge rule $b$
15: **end while**
16: $g'' := \iota(\langle g'_B, g'_D, g'_I, g'_N \rangle)$ where $\iota$ as in Definition 12 derived from $b$
17: $g := \varsigma(g'')$ where $\varsigma$ as in Definition 13
18: **for** all $(\mathcal{A}_i, V \setminus \mathcal{A}_i)$, $\mathcal{A}_i \subseteq V$ **do**
19: calculate $\sigma(g, \mathcal{A}_i)$ using Definition 4.1
20: **end for**
21: $\kappa := \kappa(g)$ using Definition 4.2
22: $\mathcal{A} := \mathcal{A}_i | \max(\sigma(g, \mathcal{A}_i))$
23: **if** $n \in \mathcal{A}$ **then**
24: $adopt$ or $comply$ norm $n$
25: **else**
26: $reject$ or $violate$ norm $n$
27: **end if**

The lines from 1 to 9 updates the graphs by incorporating the new belief and its influences on other beliefs. Lines from 10 to 15 extends the graphs of other cognitions and norms by taking into account the influence of the new belief $(B\varphi, r)$. The graph extension function $\varepsilon$ is used to extend the graphs. Note that we apply the extension function until we reach the fixed point. In line 16, the composite graph $g''$ is constructed using the graph join function $\iota$. And finally in Line 17 the composition function makes the union of all join functions to get the composite graph $g$.

In the second part of the algorithm, lines from 17 to 21 determines the coherence maximising partition. This is done by first computing the strength of each partition using the function $\sigma$ and choosing the partition $(\mathcal{A}, V \setminus \mathcal{A})$ for which $\sigma(g, \mathcal{A})$ is maximal. In the above algorithm we make an assumption that there is only one accepted set corresponding to the maximum value of $\sigma(g, \mathcal{A}_i)$. However this is not true, because even if there is just one partition $(\mathcal{A}, V \setminus \mathcal{A})$ corresponding to the maximum, $\mathcal{A}$ and its dual $V \setminus \mathcal{A}$ are already two accepted sets which produces the same maximum value of coherence. In addition there could be other partitions which maximises the coherence. According to the guidelines discussed in Chapter 4, we can decide on the 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.

Finally lines 22 to 26 checks whether the norm to be evaluated $n$ is part of the accepted set $\mathcal{A}$ associated with the coherence maximising partition. Note that this algorithm assumes that our agents are coherence maximising agents and the only factor influencing decisions is coherence maximisation. However, needless to point out that there could be other factors influencing such a decision which is not the focus of this

paper.

Another important observation is regarding the values of function $\sigma$. In theory coherence of the graph $\kappa(g)$ is set as the maximum of the strength values $\sigma(\mathcal{A}_i, V \setminus \mathcal{A}_i)$, in reality this could be very much dependent on the agent. If the inclusion of a norm only slightly reduces the coherence of the graph, a mildly distressed agent may choose to ignore the incoherence, may be satisfied with lowering the degree associated with a particular belief, may still choose to follow a norm. Where as a heavily distressed agent may not only chose to violate a norm, but initiate a dialogue to campaign for a norm change.

# Chapter 7

# Example — Norm Evaluation

We apply the formalism developed in the previous sections to model norm evaluation in a real scenario. The example is motivated by the water sharing treaty signed between the southern states of India during $1892$ and $1924$ and the disputes thereafter [29]. The objectives of this example are threefold. First, to demonstrate how self-interested agents working together evaluate norms. Second, to show the need for *norm adaptation* inspired by individual coherence evaluations, whereas the grander aim is to set up a framework for norm adaptation itself, which will be our future work. And third, to open new application areas in norm evaluation where such cognitive theories could be applied.

We describe now the reasoning performed by a coherence-driven agent. We simplify the case for brevity, considering just two agents $s$ and $t$ standing for two distinct Indian states. We model the reasoning of $s$ in two snapshots of time $t_1$ and $t_2$, one when the first treaty is about to be signed (i.e, the decision to adopt the norm) and the second after a period of working together, when the situation has evolved.

## 7.1 Terminology

To represent the cognitions and norms concerning an agent, we shall have belief, desire, intention and norm languages as defined in Section 6.1. Hence, $(B\varphi, r)$ represents that the agent believes that proposition $\varphi$ is true (in a near future world[1]) with degree $r$. (Propositions $(D\varphi, r)$, and $(I\varphi, r)$ are desires and intentions and are interpreted analogously). The statements about the world are in propositional language where each proposition is a triple of the form $\langle object, attribute, value\rangle$. For instance $\langle urbanization, growth\_index, high\rangle$ states that *there is a high growth in urbanisation*. Note that, we do not have an explicit norm coherence graph in this example, as graded normative language is under research currently. However, we represent norms accepting a norm by beliefs about the norms and intentions on its consequences.

[1] In our representation we refer to future worlds as the agent is trying to anticipate the coherence of future worlds where the norm is accepted or rejected.

| | |
|---|---|
| $\varphi_{11}$ | $\langle river\_basin, water\_index, adequate\rangle$ |
| $\varphi_{12}$ | $\langle rain\_fall, index, good\rangle$ |
| $\varphi_{13}$ | $\langle water\_release, quantity,$ *300 billion ft*$^3\rangle$ |
| $\varphi_{14}$ | $\langle s_2\_threat, type, military\_force\rangle$ |
| $\varphi_{15}$ | $\langle s_2\_threat, status, realised\rangle$ |
| $\varphi_{16}$ | $\langle treaty\_proposal, status, accepted\rangle$ |
| $\varphi_{17}$ | $\langle internal\_demand, status, satisfied\rangle$ |

Table 7.1: Propositions relevant for $s_1$'s cognitions at $t_1$

The bridge rules we use in the water-sharing example are the following. These are chosen for illustration purposes, however the bridge rules can be chosen according to the characteristics of the agent we want to model.

1. $b_1 = \frac{3:(I\psi,r)}{2:(B\psi,r)}$: Whenever there is an intention $(I\psi, r)$ in context $C_3$, then a corresponding belief $(B\psi, r)$ is inferred in context $C_1$.

2. $b_2 = \frac{1:(B\psi,r),2:(D\psi,s)}{3:(I\psi,\min(r,s))}$: Whenever there is a belief $\psi$ with a degree $r$ in $C_1$ and a desire $\psi$ with a degree $s$ in $C_2$, then a corresponding intention $\psi$ with a degree $\min(r, s)$ is inferred in the context $C_3$.

## 7.2 Norm Adoption

*Snapshot* $t_1$ : 1892
*Agent :* $s$
*Action*: Evaluating the proposal of the water sharing treaty.
*Norm to be evaluated*: agent $s$ should release 300 billion ft$^3$ of water to agent $t$ annually.

Agent $s$ reasons by injecting into its internal coherence graph $g_1 = \langle V_1, E_1, \rho_1, \zeta_1\rangle$, the anticipated consequences of the norm adoption and compares the coherence on signing the treaty as opposed to not signing it. Here we use coherence as the primary mechanism for decision making, however, in the future we shall analyse also the influence of sanctions and rewards. Although in our framework sanctions related to norms are not modeled explicitly, we take into account their influences in forming the agent modalities. Below we list the propositions relevant to forming the agent cognitions and then the cognitions $V_1$ of agent $s$ at $t_1$:

Below we analyse the hypothetical reasoning that the agent $s$ does to evaluate the norm, i.e $(\varphi_{16} \rightarrow O\varphi_{13}, 1)$ , where $\varphi_{16} = \langle treaty\_proposal, status, accepted\rangle$ and $\varphi_{13} = \langle water\_release, quantity,$ *300 billion ft*$^3\rangle$. Since we reason in terms of the beliefs about the norms, we have the following to represent the norm:

| | | |
|---|---|---|
| $Beliefs$ | $(B\varphi_{11}, 0.90)$ | $(B\varphi_{12}, 0.75)$ |
| | $(B\varphi_{14}, 1)$ | |
| | $(B\varphi_{11} \wedge \varphi_{12} \wedge \varphi_{13}) \rightarrow \varphi_{17}, 1)$ | |
| | $(B\varphi_{14} \wedge \neg\varphi_{16} \rightarrow \varphi_{15}, 1)$ | $(B\varphi_{16} \rightarrow \neg\varphi_{15}, 1)$ |
| $Desires$ | $(D\varphi_{17}, 0.95)$ | $(D\neg\varphi_{15}, 1)$ |
| $Intentions$ | $(I\varphi_{17}, 0.95)$ | $(I\neg\varphi_{15}, 1)$ |

Table 7.2: $s_1$'s cognitions ($V_1$) at snapshot $t_1$

### 7.2.1 Case 1: $s$ accepts the treaty $\{B(\varphi_{16}, 1), (I\varphi_{13}, 1)\}$

Accepting to sign the treaty is equivalent to incorporating an additional belief that at a near future world, ($\varphi_{16}$ is true with degree 1 and an additional intention to make it true. That is, $V_1 := V_1 \cup \{(B\varphi_{16}, 1), (I\varphi_{13}, 1)\}$. Below we calculate the coherence of the agent in conjunction with these additional cognitions. Applying the *max-cut* algorithm, we have the coherence maximising partition $(\mathcal{A}, V \setminus \mathcal{A})$ as shown in the Figure 7.1. The corresponding coherence of the graph, $\kappa(g_1)$ is $4.41/16 = 0.275625$.

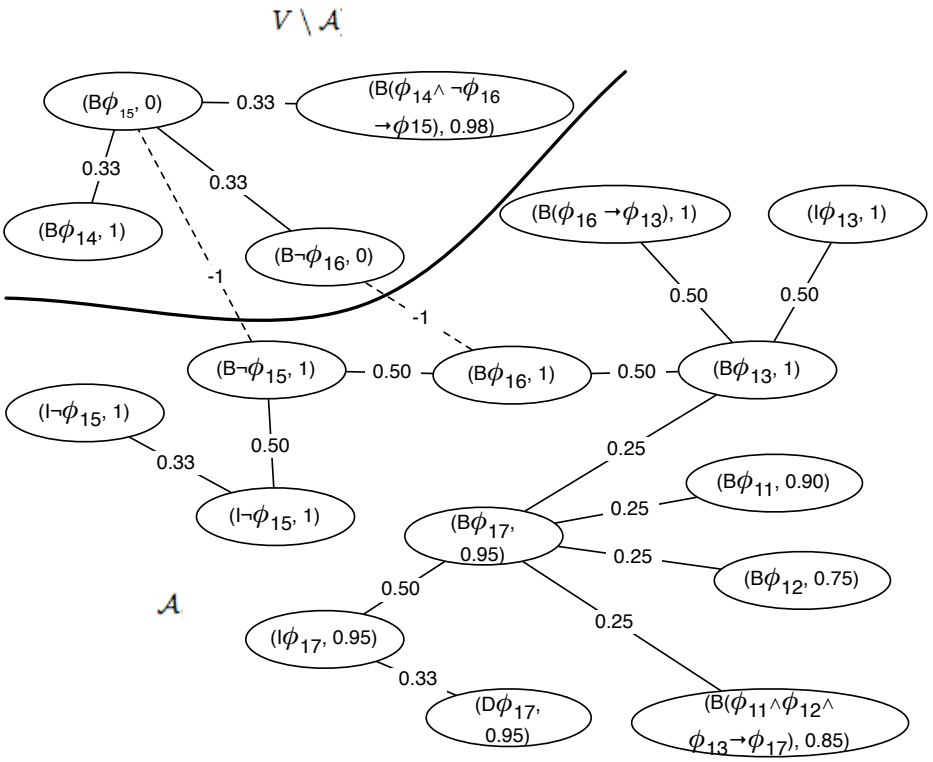


Figure 7.1: Coherence graph ($g_1$), with norm accepted $\kappa(g_1) = 0.275625$

| | |
|---|---|
| $\varphi_{21}$ | $\langle urbanisation, growth\_index, high\rangle$ |
| $\varphi_{22}$ | $\langle industrialisation, growth\_index, high\rangle$ |
| $\varphi_{23}$ | $\langle water\_usage, growth\_index, high\rangle$ |
| $\varphi_{24}$ | $\langle revenue, growth\_index, high\rangle$ |

Table 7.3: Propositions related to $s_1$'s cognitions at snapshot $t_2$

| | | |
|---|---|---|
| $Beliefs$ | $(B\neg\varphi_{11}, 0.2)$ | $(B\neg\varphi_{12}, 0.75)$ |
| | $(B\varphi_{14}, 0.75)$ | $(B\varphi_{21}, 0.90)$ |
| | $(B\varphi_{22}, 0.90)$ | $(B\varphi_{13}, 1)$ |
| | $(B\neg\varphi_{11} \wedge \neg\varphi_{12} \wedge p_{23} \wedge \varphi_{13} \rightarrow \neg\varphi_{17}, 0.90)$ | $(B\varphi_{14} \wedge \neg\varphi_{16} \rightarrow \varphi_{15}, 0.75)$ |
| | $(B\varphi_{21} \wedge \varphi_{22} \rightarrow \varphi_{23}, 1)$ | $(B\varphi_{24} \rightarrow \varphi_{21}, 1)$ |
| | $(B\varphi_{24} \rightarrow \varphi_{22}, 1)$ | $(B\varphi_{17} \rightarrow \varphi_{24}, 0.75)$ |
| | $(B\varphi_{16} \rightarrow \neg\varphi_{15}, 1)$ | $(B\varphi_{24} \rightarrow \varphi_{23}, 0.80)$ |
| $Desires$ | $(D\varphi_{17}, 0.95)$ | $(D\varphi_{24}, 0.85)$ |
| | $(D\neg\varphi_{15}, 1)$ | |
| $Intentions$ | $(I\varphi_{17}, 0.95)$ | $(I\varphi_{24}, 0.85)$ |
| | $(I\neg\varphi_{15}, 1)$ | $(I\varphi_{16}, 1)$ |

Table 7.4: $s_1$'s cognitions at snapshot $t_2$

### 7.2.2 Case 2: $s$ rejects the treaty $(B\neg\varphi_{16}, 1)$

The differences if $s$ decides not to accept the treaty are that it has the additional belief $(B\varphi_{15}, 1)$ whereas it removes the intention $(I\neg\varphi_{15}, 1)$ as it is reasonable to assume that agent $t$ will realise the threat upon rejecting the treaty. That is $V_1 := V_1 \cup \{(B\neg\varphi_{16}, 1), (B\varphi_{15}, 1)\} \setminus \{(I\neg\varphi_{15}, 1)\}$. With these changes, we have the the coherence of the graph as $\kappa(g_1) = 3.07/16 = 0.191875$. As a coherence-driven agent seeks coherence maximisation, $s$ prefers to adopt the treaty guided by its coherence value. However we do not rule out the possibilities of other considerations of the agent that can influence its final decision.

## 7.3 The Incoherence Buildup

*Snapshot* $t_2$ : 1991
*Agent* : $s$
*Action*: Updating cognitive graph based on situation change.
*New Facts*: $s$ experiences large-scale industrialisation, urbanisation, higher water usage, threat from $t$ to obey the norm, and less amount of rain fall.
Below we list the propositions capturing this change in situation and the changed cognitions of the agent $s$ at $t_2$:

The coherence graph $g_2$ of the agent $s$ with changed cognitions is shown in Figure 7.2. Some of the cognitions that do not influence the result have not been included in $g_2$ for the sake of clarity. Using the coherence equations, the coherence maximis-

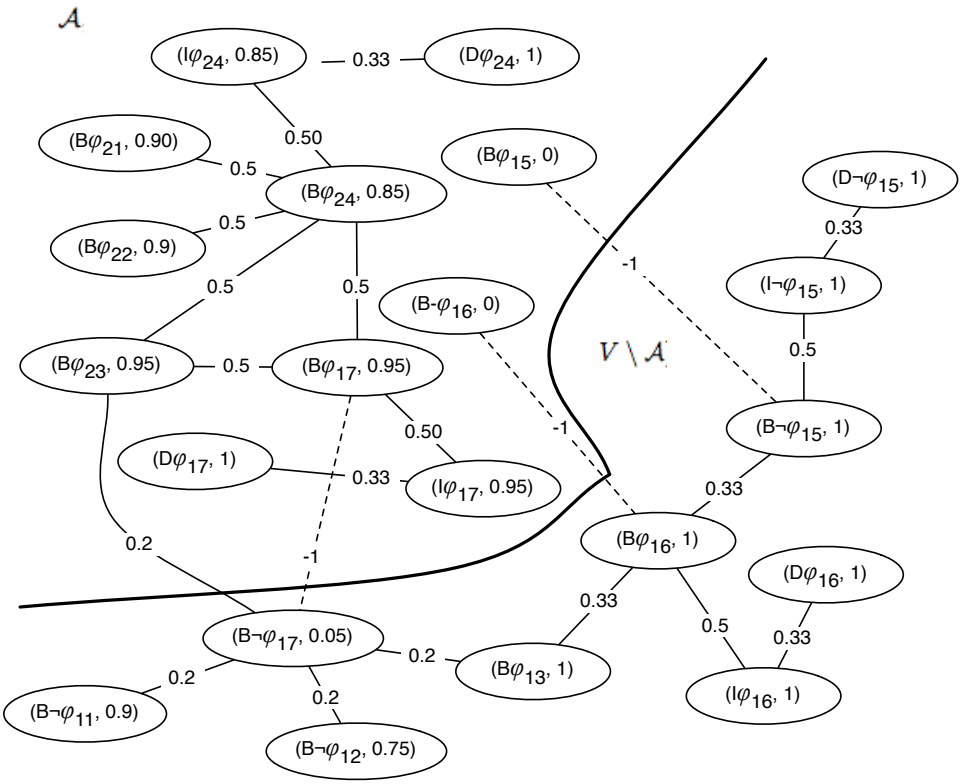


Figure 7.2: Coherence graph ($g_2$), $\kappa(g_2) = .29$

ing partition $(\mathcal{A}, V \setminus \mathcal{A})$ is shown in Figure 7.2. The partition interestingly places the cognitions about $\varphi_{24}$ and $\varphi_{17}$ in set $\mathcal{A}$ while the cognitions about $\varphi_{16}$ and $\neg\varphi_{15}$ in set $V \setminus \mathcal{A}$. It is clear from the coherence evaluation that all these intentions cannot coexist while maintaining the maximum coherence. That is, the agent has to choose between *obeying the norm, hence avoiding the threat of military action* and *satisfying the internal demands for water, hence economic progress*. Even though the ultimate decision can vary from other considerations of the agent, a purely coherence maximising agent will choose to violate the norm in order to keep a maximal state of coherence. With this example we show how a coherence maximising agent evaluates norms in the context of its cognitions.

## 7.4 Discussion

Even though the example only demonstrates the case of a single norm, the same can be extended to cases where there are multiple norms and there is a need to choose among the norms. In terms of coherence, this is selecting a norm (or a set of norms) which

maximises the coherence of the graph. By performing the hypothetical analysis of a norm being accepted, norms can be ordered according to the coherence each would generate in the resulting adoption. Another point to note is that here we have assumed our agents to be coherence maximising. But in reality there are other criteria that need to be considered. Some of them already mentioned and represented in the graph are sanctions and rewards. Another important factor by which an agent makes a decision to adopt a norm is observing the behavior of other agents. We can represent this as cognitive models of other agents.

### 7.4.1 Computational Complexity

We have implemented some important functions of the coherence framework. We can show that the popular max-cut problem can be converted to an equivalent coherence maximising problem. As max-cut is an NP-complete problem it becomes clear that coherence maximisation is also an NP-complete problem. However neural network based algorithms give good approaximation. Thagard in his formalisation of coherence has given several implementations of coherence , with an extensive implementation of a neural network model called ECHO [26]. He also compares it with a max-cut implementation. We have extended Thagard's implementation to incorporate additional features of our model. That is, The solutions in our implementation we use a Prolog-based meta interpreter to extract proofs of each sentence in the BDI base of the agent where these proofs will give raise to the coherence values between pairs of sentences using the support function $\eta$ of Section 3. We further use a semi-definite programming max-cut approximation algorithm to evaluate the coherence of the graph and to determine the nodes in the accepted set [28].

# Chapter 8

# Conclusions and Ongoing Work

In this dissertation, we proposed a coherence based framework which extends the popular BDI architecture by including the notion of coherence. Coherence-driven agents take actions based on coherence maximisation as opposed to utility based agents maximising utility. We show that coherence maximisation gives the necessary autonomy to the agents to take decisions considering dependencies among cognitions and external commitments. We show, in particular, how an agent could evaluate a norm by maximising coherence. We provide a coherence function and its properties to construct coherence graphs from a set of pieces of information. We also provide functions to compute the coherence of a graph and for composing different types of graphs, so that, an agent could consider the overall effects of different cognitions and external commitments.

However there are questions asked about the philosophy behind coherence. One question often put forward with respect to the application of coherence as an agent decision making tool is whether it is rational for an agent to behave according to coherence maximisation. Normally an agent reasoning about norms takes into account influences of utility maximisation, models of other agents, and sanctions or rewards. We claim that we can introduce these decision making factors into our coherence graph so that the coherence maximisation is the only evaluation necessary for the decision making process. To demonstrate that coherence can actually contain these factors, in this chapter we show how utility maximisation (in the terminology of game theory) can be reduced to coherence maximisation.

In fact, we see that coherence can be used to model different types of agents, a utility maximising rational agent, a norm abiding institutional agent, a selfish agent, or more importantly an autonomous social agent. We show in the following section that a utility function can be reduced to a specific type of coherence graph. For the norm abiding institutional agents, a coherence maximisation immediately provides a clear answer. If there are conflicts between norms and cognitions, the coherence maximising partition puts the norms on one set of the partition, whereas the cognitions go to its complimentary set. Then a norm abiding agent always chose the set with the norms included. A selfish agent on the other hand, always chose the set with its cognitions included ignoring the norms.

Another important question is about the credibility of a coherence-based evaluation. For example, coherence-based evaluation gives an impression of a black box churning out results, when compared to, for instance, an argumentation-based system, which would give explicit reasons why an argument is the best in terms of its de-feasibility. The difference between an argumentation system and a coherence-based system is that, a coherence-based system maximises the overall coherence, giving less importance to particular pieces of information, where as in an argumentation system, its a particular piece of information that is important than the overall system. Further, we have based our framework on sound theoretic framework, computing the coherence graph using the properties of the coherence function.

Another related question is about the computational feasibility of coherence maximisation. Unlike other proposals on coherence maximisation, in this dissertation, we have introduced a fully computational framework of coherence. However, as we have stated in Section 7.4.1, coherence maximisation is an NP-complete problem. But, as we assume bounded rationality for our agents, our coherence graphs are also bounded and is not complete. Further we are exploring ways to bring in contexts, which would consider only a sub-graph of the actual with the intuition that coherence maximisation should consider only those nodes which are relevant to the problem at hand. In Section 8.2, we explore the semantic interpretation of coherence, which when extended gives us the notion of an evidential set. We plan to explore this as part of our future work.

In the remaining sections we list some of the directions we are working on. These are intended to give a glimpse of the general direction of research, however, not perfected from a technical point of view.

## 8.1 Coherence as an Inclusive Notion of Rationality

As stated in the introduction to the chapter, we here take up a specific notion in game theory and show how that notion can be reduced to a coherence graph. The aim of this reduction is to demonstrate that, the notion of utility which is fundamental to a rational agent can be modelled in a coherence graph. We however assume that there is an apriori preference established by the agent.

Preferences are relevant when it is necessary to examine the behaviour of an individual, called a player, who must choose from among a set of outcomes $O$.

**Definition 17** *A preference relation $\succeq$ on $O$ is a binary relation on $O$ such that $\forall o, p \in O$*

1. *$o \succeq p$ if and only if $o$ is at least as preferable as $p$.*
2. *if $o \succeq p$ but not $p \succeq o$, then $o$ is strictly prefered $p$ denoted as $o \succ p$.*
3. *if $o \succeq p$ and $p \succeq o$ then $o$ is equivalent to $p$ denoted as $o \sim p$.*

We assume that the following properties hold for our strict preference relation $\succ$ and equivalence relation $\sim$. These properties can be derived from the more basic assumptions that $\succ$ is asymmetric and transitive. $\forall o, p, s, t \in O$

1. $\succ$ is complete: if $o \neq p$, then either $o \succ p$ or $p \succ o$ or both
2. $\succ$ is transitive: if $o \succ p$ and $p \succ s$, then $o \succ s$
3. $\sim$ is reflexive: $o \sim o$
4. $\sim$ is symmetric: if $o \sim p$ then $p \sim o$
5. $\sim$ is transitive: if $o \sim p$ and $p \sim s$, then $o \sim s$
6. if $t \sim o$, $o \succ p$, and $p \sim s$, then $t \succ p$ and $o \succ s$.

A utility function assigns a numerical value to each of the outcomes in $O$ such that the preference relation is respected. There are many utility functions that can represent the same preference relation. Note that preferences are ordinal, that is, they specify the ranking of the alternatives, but not how far apart they are from each other (intensity). Cardinal properties are those that are not preserved under strictly increasing transformations. For example, because the function assigns numerical values to various alternatives, the magnitude of any differences in the utility between two alternatives is cardinal. We want to use a utility function because instead of examining conditions under which preference relations produce maximal elements for a set of alternatives, it is easier to specify the numerical representation and then apply standard optimisation techniques to find the maximum. Thus, the best options from the set $O$ are precisely the options that have the maximum utility. More formally,

**Definition 18** *A function $u : O \rightarrow R$ is a* utility function representing preference relation $\succ$ *if the following holds for all $o, p \in O$: $o \succ p \Leftrightarrow u(o) > u(p)$.*

At this point, it is worth emphasising that players do not have utility functions. Rather, they have preferences, which we can represent (for analytical purposes) with utility functions.

### 8.1.1 A Utility Coherence Graph

Now we reduce the utility function for preference relation $\succ$ to a coherence graph with certain properties. Our aim in this reduction is to show that such a graph can be constructed for any preference relation $\succ$ which satisfies the properties in 8.1.

**Theorem 8.1.1** *Given a finite set of outcomes $O$ and a utility function $u : O \rightarrow [0, 1]$ there exists a coherence graph $g$ such that the accepted set $\mathcal{A}$ of its coherence maximising partition is*

$$\mathcal{A} = \{\arg\max_{o \in O} u(o)\}$$

To prove the theorem, we need to prove the following propositions.

**Proposition 8.1.2** *Given a finite set of outcomes $O$ and a utility function $u : O \rightarrow [0, 1]$, there exists a coherence graph $g$ such that*

1. $\forall o, p \in O$, $u(o) > u(p) \Leftrightarrow \sigma(g, \{o\}) > \sigma(g, \{p\})$

2. *Given* $O' \subseteq O$ *such that* $0 \leq |O'| \leq \lfloor \frac{|O|}{2} \rfloor$ *and* $|O'| \neq 1$ *we have*

$$\forall o \in O' \quad \sigma(g, \{o\}) > \sigma(g, O')$$

**Sketch of the proof:** Let $g = \langle O, E, \rho, \zeta \rangle$ be a coherence graph such that

- $E$ are all subsets of 2 elements of $O$
- for all $o, p \in O$ $\quad u(o) > u(p)$ iff $\rho(o) > \rho(p)$
- for all $o \in O$ $\quad \rho(o) > \max_{S \subseteq O \setminus \{o\}, |S| = n-2} \Sigma_{q \in S} \rho(q)$
- for all $o, p \in O$ $\quad \zeta(o, p) = -1$

**1.:**

$$\begin{aligned} u(o) > u(p) \quad &\Leftrightarrow \quad \rho(o) > \rho(p) \text{ (property of } g) \\ &\equiv \quad \rho(o) \cdot \Sigma_{q \in O \setminus \{o,p\}} \rho(q) + \rho(o) \cdot \rho(p) > \rho(p) \cdot \Sigma_{q \in O \setminus \{o,p\}} \rho(q) \\ &\quad\; + \rho(p) \cdot \rho(o) \\ &\equiv \quad \rho(o) \cdot \Sigma_{q \in O \setminus \{o\}} \rho(q) > \rho(p) \cdot \Sigma_{q \in O \setminus \{p\}} \rho(q) \\ &\equiv \quad \sigma(g, \{o\}) > \sigma(g, \{p\}) \end{aligned}$$

**2.:**
Let $O' \subseteq O$ such that $0 \leq |O'| \leq \lfloor \frac{|O|}{2} \rfloor$ and $|O'| \neq 1$.
Then we have that

$$\max_{S \subseteq O \setminus \{o\}, |S| = n-2} \Sigma_{q \in S} \rho(q) \quad > \quad \Sigma_{q \in O \setminus O'} \rho(q)$$

Consequently,

$$\begin{aligned} \rho(o) &> \Sigma_{q \in O \setminus O'} \rho(q) \\ \rho(o)(\Sigma_{q \in O' \setminus \{o\}} \rho(q) + \Sigma_{q \in O \setminus O'} \rho(q)) &> (\rho(o) + \Sigma_{q \in O' \setminus \{o\}} \rho(q)) \Sigma_{q \in O \setminus O'} \rho(q) \\ \rho(o) \cdot \Sigma_{q \in O \setminus \{o\}} \rho(q) &> \Sigma_{q \in O'} \rho(q) \cdot \Sigma_{q \in O \setminus O'} \rho(q) \\ \sigma(g, \{o\}) &> \sigma(g, O') \end{aligned}$$

Hence, we may prove the theorem as follows:

1. From Proposition 1.1 we have that given $u(o) > u(p)$, coherence maximisation would always select $\{o\}$ over $\{p\}$ as the accepted set since $\sigma(g, \{o\}) > \sigma(g, \{p\})$.
2. Proposition 1.2 states that, the value of $\sigma$ for a singleton set is always higher that that of any set which includes this singleton set. Hence coherence maximisation always select singleton sets as accepted sets over any set including them.

However, note that a graph with the stated properties is not easy to imagine. Our future work includes restating the properties purely in terms of function $\rho$ or in other words, finding a distribution for values of $\rho$ which satisfies the stated properties of the graph.

## 8.2 A Semantical Approach

Using proof-theory, we have demonstrated that coherence is a fully computable phenomena. However, to expand its reach to all types of coherence, we need a more general definition. Here we introduce the semantic interpretation of coherence in terms of degrees of consistency by Ruspini. That is, we interpret coherence as a similarity function between possible world. This interpretation gives us a larger framework to analyse different types of coherence by changing the similarity function.

In this section we propose a semantical formalisation of coherence using the notion of *degrees of consistency* introduced by Ruspini in [23]. Ruspini in his work interprets the similarity between two propositions, by the similarity between the worlds in which the propositions are true. Using this interpretation, we define coherence as the similarity between possible worlds.

We first introduce the basic definitions from Ruspini's degree's of consistency and then define coherence in terms of it. For the sake of clarity we restrict now our attention to propositional languages. Let $L$ be a propositional language and $W$ a set of classical interpretations of $L$ (i.e. a set of possible worlds). For any $w \in W$ and any proposition $p \in L$, we denote by $w \models p$ the fact that proposition $p$ is true in the interpretation $w$. First we introduce some basic definitions.

**Definition 19** *A function* $T : [0,1] \times [0,1] \to [0,1]$ *is a* triangular norm *if and only if:*

1. $T$ *is commutative and associative*
2. $T$ *is non-decreasing in both arguments*
3. $T(1,x) = x$ *and* $T(0,x) = 0$ *for all* $x \in [0,1]$

**Definition 20** *Given a triangular norm* $T$, $S_T : W \times W \to [0,1]$, *where* $T$ *is a triangular norm where the function is continuous (t-norm for short), is a* T-similarity function *if and only if* $S_T$ *satisfies the following properties:*

1. *Reflexivity:* $S_T(w,w) = 1$
2. *Symmetry:* $S_T(w,w') = S_T(w',w)$
3. *T-Transitivity:* $S_T(w,w') \geq T(S_T(w,w''), S_T(w'',w'))$

The function assigns a degree of similarity between 0 (corresponding to maximum dissimilarity) and 1 (corresponding to maximum similarity). For the sake of simplicity, $S_T$ is required to fulfill that if$S_T(w,w') = 1$ then $w = w'$. The transitivity requirement allows $S_T$ to become a generalised equivalence relation.

Ruspini generalises the semantical entailment relationship between propositions. He defines both an implication function and a consistency function between propositions. The definition of partial implication between propositions is based on conditions that determine whether, given two propositions $p$ and $q$, one of them implies the other to the degree $n$. We introduce the formal definitions below:

**Definition 21** *Given a T-similarity relation $S_T$ and propositions $p, q \in L$, the* degree of implication $Imp(p \mid q)$ *is given by :*

$$Imp(p \mid q) = \inf_{w' \models q} \sup_{w \models p} S_T(w, w')$$

*Given a T-similarity relation $S_T$ and propositions $p, q \in L$, degree of consistency $Con(p \mid q)$ is given by:*

$$Con(p \mid q) = \sup_{w' \models q} \sup_{w \models p} S_T(w, w')$$

Observe that the degree of consistency $Con$ is a symmetric measure while the degree of implication $Imp$ is not. Nevertheless, $Imp$ has the T-transitivity property of similarity. Moreover, for any formulas $p, q \in L$, $Con(p \mid q) \geq Imp(p \mid q)$.

By definition of the implication and consistency measures it is easy to check that $Imp(p \mid q) = 1$ iff $q \models p$ whereas $Con(p \mid q) = 1$ iff $q \not\models \neg p$. Now we state some basic properties of the consistency degree for $L$ with $\delta, \beta, \gamma \in L$ and $n, m \in [0, 1]$:

1. $Con(\delta \wedge \beta \mid \beta) = 1$ iff $\delta, \beta \not\models$
2. $Con(\delta \vee \beta \mid \beta) = 1$ iff $\delta, \beta \not\models$
3. If $Con(\gamma \mid \delta) = n$ and $Con(\gamma \mid \beta) = m$, then $Con(\gamma \mid \delta \vee \beta) = \max(n, m)$
4. If $Con(\gamma \mid \delta \wedge \beta) = 1$ then $Con(\gamma \mid \delta) = 1$ and $Con(\gamma \mid \beta) = 1$
5. If $Con(\gamma \mid \delta \vee \beta) = n$ then $Con(\gamma \mid \delta) \leq n$ and $Con(\gamma \mid \beta) \leq n$
6. $Con(\delta \mid \neg\delta) = 0$

Now we can define a *coherence function* $\kappa' : \mathcal{T} \times \mathcal{T} \to [-1, 1]$ on $\mathcal{T}$ in terms of degrees of consistency as follows:

**Definition 22** *Let $\mathcal{T}$ be a finite set of formulas of a language L. For any pair $(\delta, \beta)$ of formulas in $\mathcal{T}$, a* coherence function $\kappa' : \mathcal{T} \times \mathcal{T} \to [-1, 1]$ *on $\mathcal{T}$ is*

$$\kappa'(\delta, \beta) = \begin{cases} undefined & \text{if } Con(\delta, \beta) = 0 \\ 2 \cdot Con(\delta, \beta) - 1 & otherwise \end{cases}$$

Note that, the coherence function $\kappa$ always takes nonzero values. However, he consistency function $Con$ is defined for the entire interval $[-1, 1]$. If $Con(\delta, \beta) = 0$ for any propositions $\delta$ and $\beta$, then there is no consistency between worlds in which $\delta$ is true and those in which $\beta$ is true. Hence, we can safely make $\kappa(\delta, \beta)$ as $undefined$.

## 8.3 Future Work

In this dissertation, we have demonstrated a coherence-driven agent, or in other words, we have shown how a coherence based framework is beneficial to an individual agent. Coherence not only is interesting to an individual agent, but is also a relevant concept at

the social level. In our future work, we would like to take coherence to the social level, where parallel concepts of Pareto efficiency and Nash equilibrium can be defined in terms of coherence. This would give the agents a more general notion than rationality, which encompasses notions of fairness. Thus, our aim is to provide a more general framework for normative multiagent systems which is dynamic and have equilibrium conditions in terms of coherence.

In this dissertation, we have formalised one specific type of coherence, namely the deductive coherence. However, deductive coherence was chosen, only for the sound theoretical base that we have in the underlying logic. However for building practical coherence-driven agents, we need to look at other types of coherence. With the help of ideas introduced in Section 8.2 of going work on semantical interpretation, we would like to ex tend formalisation of coherence which would include many other types. Since the semantic interpretation of coherence depends on a similarity function, it is possible to define this function between propositional elements in the case of deductive coherence, or between concepts in the case of explanatory coherence. This would help provide a more holistic notion of coherence.

In this dissertation, the treatment of norms were inadequate. More precisely, we did not represent norms directly, but accepting a norm was equivalent to a belief in the antecedent part of the norm and an intention to realise its consequent. That is we expressed norms in terms of cognitions. However, we plan to have a norm language similar to that of graded belief languages. As research on probabilistic standard deontic logic [10] progresses, we incorporate this language into our framework. One of our ongoing work, not mentioned here is to make the concept of norm more precise in the language of coherence. Following many others, we treat norms as constraints on cognitions, and we go one step forward by considering them as coherence constraints between cognitions. This is one of our immediate future work.

We also would like to compare similar systems with our coherence framework such as argumentation systems from practical reasoning, legal reasoning domains, constraint satisfaction models from optimisation, dominant principles from economics.

# Bibliography


[1] Amalia Amaya. Formal models of coherence and legal epistemology. *Artif. Intell. Law*, 15(4):429–447, 2007.

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

[3] Guido Boella and Leendert van der Torre. Fulfilling or violating obligations in normative multiagent systems. In *IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'04)*, 2004.

[4] Guido Boella, Leendert van der Torre, and Harko Verhagen. Introduction to normative multiagent systems. In *Normative Multi-agent Systems*, 2007.

[5] Michael E. Bratman. *Intention, Plans, and Practical Reason*. CSLI publications, 1987.

[6] Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *AGENTS '01*, 2001.

[7] Ana Casali, Llus Godo, and Carles Sierra. A methodology to engineer graded bdi agents. In *WASI - CACIC Workshop.XII Congreso Argentino de Ciencias de la Computacin*, 2006.

[8] Cristiano Castelfranchi, Frank Dignum, Catholijn M. Jonker, and Jan Treur. Deliberative normative agents: Principles and architecture. In *ATAL '99*, 2000.

[9] Rosaria Conte, Cristiano Castelfranchi, and Frank Dignum. Autonomous norm acceptance. In *ATAL '98*. Springer-Verlag, 1998.

[10] Pilar Dellunde and Lluis Godo. *Introducing Grades in Deontic Logics*. LNAI, Springer, to appear., 2008.

[11] Paul E. Dunne and T J. M. Bench-Capon. Coherence in finite argument systems. *Artif. Intell.*, 141(1):187–203, 2002.

[12] K. Brad Wray (ed.). *Knowledge and Inquiry*. Broadview Press, 2002.

[13] David Fitoussi and Moshe Tennenholtz. Choosing social laws for multi-agent systems: minimality and simplicity. *Artif. Intell.*, 119(1-2):61–101, 2000.

[14] Fausto Giunchiglia and Fausto Giunchiglia. Contextual reasoning. *Epistemologia, special issue on I Linguaggi e le Macchine*, 345:345–364, 1993.

[15] Fausto Giunchiglia and Luciano Serafini. Multilanguage hierarchical logics, or: how we can do without modal logics. *Artif. Intell.*, 65(1):29–70, 1994.

[16] P. Hájek. Metamathematics of fuzzy logic. In *Trends in Logic*, volume 4, 1998.

[17] Martin J. Kollingbaum and Timothy J. Norman. 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, 2003.

[18] Paul K. Moser(ed.). *The oxford handbook of epistemology*. Oxford university press, 2002.

[19] Yoram Moses and Moshe Tennenholtz. Artificial social systems. *Computers and AI*, 14:533–562, 1995.

[20] Philippe Pasquier and Brahim Chaib-draa. The cognitive coherence approach for agent communication pragmatics. In *AAMAS '03*, 2003.

[21] Piwek. Meaning and dialogue coherence: a proof-theoretic investigation. *Journal of Logic, Language and Information*, 16(4), 2007.

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

[23] Enrique H. Ruspini. On the semantics of fuzzy logic. *International Journal of Approximate Reasoning*, 5, 1991.

[24] Jean-Paul Sansonnet and Erika Valencia. A model for dialog between semantically heterogeneous informational agents. In *EPIA '03, MAAII*, 2003.

[25] Yoav Shoham and Moshe Tennenholtz. On social laws for artificial agent societies: off-line design. *Artif. Intell.*, 73(1-2):231–252, 1995.

[26] Paul Thagard. *Coherence in Thought and Action*. MIT Press, 2002.

[27] Paul Thagard. *Hot Thought*. MIT Press, 2006.

[28] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Rev.*, 38, 1996.

[29] Wikipedia. Kaveri river water dispute — wikipedia, the free encyclopedia, 2008.

[30] Michael Wooldridge. *Reasoning about rational agents*. MIT press., 2000.

[31] Fabiola López y López, Michael Luck, and Mark d'Inverno. Constraining autonomy through norms. In *AAMAS '02*, 2002.