Logic Journal of IGPL Advance Access published December 16, 2009

Deductive coherence and norm adoption

SINDHU JOSEPH, *Artificial Intelligence Research Institute, IIIA, CSIC, Bellaterra (Barcelona), Catalonia, Spain.*
E-mail: joseph@iiia.csic.es

CARLES SIERRA, *Artificial Intelligence Research Institute,IIIA, CSIC, Bellaterra (Barcelona), Catalonia, Spain.*
E-mail: sierra@iiia.csic.es

MARCO SCHORLEMMER, *Artificial Intelligence Research Institute, IIIA, CSIC, Bellaterra (Barcelona), Catalonia, Spain.*
E-mail: marco@iiia.csic.es

PILAR DELLUNDE, *Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Catalonia, Spain.*
E-mail: pilar.dellunde@uab.cat

Abstract

This paper 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. 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. In particular, we interpret coherence maximisation as a decision-making criterion for norm adoption. We first provide a general coherence framework with the necessary computing tools. Later we introduce a proof-theoretic characterisation of a particular type of coherence, namely the deductive coherence based on Thagard's principles, and derive a mechanism to compute coherence values between elements in a deductive coherence graph. Our use of graded logic helps us to incorporate reasoning under uncertainty, which is more realistic in the context of multiagent systems. We then conduct a case study where agents deliberate about norm adoption in a multiagent system where there is competition for a common resource. We show how a coherence-maximising agent decides to violate a norm guided by its coherence.

Keywords: deductive coherence, multiagent systems, normative systems

1 Introduction

A normative multiagent system is a multiagent system where the agent interactions are governed by norms. In the context of this paper, norms are associated with regulatory norms such as obligations, permissions and prohibitions, which specify the ideal behaviour of agents. While a normative multiagent system is prescriptive about agent behaviour, it

doi:10.1093/jigpal/jzp074

does so within the framework of autonomous agents. That is, the system assumes agents to reason about norms autonomously. This is because the success of such systems does not depend on all the agents following the prescribed norms blindly, rather on having rational agents deliberate about the prescribed norms, evaluate their usefulness, selectively follow those norms that improve their efficiency, and effect a change when there are conflicts, inefficiencies, or situational changes that they can perceive. We understand such agents as norm-autonomous and refer to the term 'autonomy' in this respect.

We are certainly not the first to identify this need, and there have been numerous attempts in the recent past explicitly addressing this issue [7, 14, 23, 25, 27, 36]. Many of these efforts are focused towards extending the cognitive agent theory (for instance the Belief, Desire, and Intention theory [29]) with explicit representations of norms (BOID [9], EMIL [14], and NoA [23]). However, apart from providing static-priority based autonomy[1] and recognising autonomous norm acceptance phases, a gap still exists in creating norm-autonomous agents.

To enhance the autonomous capabilities of agents, we propose a normative agent theory which extends the BDI theory with the theory of coherence [31]. 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 maximising the constraints imposed by situations and other cognitions. Pasquier et al. [27] 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 specify how a graph representing these relations can be constructed. Thus, a general computational model of coherence and coherence-based reasoning in norm-autonomous agents is still required.

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 each other (yielding a negative constraint). 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 pieces of information, a graph representation is most intuitive. Normally a graph with nodes and weighted edges is 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 paper 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. However, these results could be applied in analysing other types of coherence. Our aim is to generate coherence values between pairs of pieces of information (in this case, formulas in a logical language) by formalising the relationship between coherence and logical

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

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 objectives aimed at in this paper are two:

1. to find a reasoning mechanism for norm-autonomous agents that enables them to resolve conflicts among cognitions and norms;
2. to determine if and how such a reasoning tool can be incorporated in an agent theory such as the BDI theory.

The first objective is, thus, to look for an appropriate theory for norm-autonomous agents. To address it, we propose the theory of coherence as the reasoning mechanism for designing norm-autonomous agents. 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. The second objective is to define a computational model for this theory. We address it by proposing a coherence-driven BDI agent architecture that incorporates a reasoning algorithm.

Thus the main contributions in this paper are:

1. A formalisation of Thagard's principles of deductive coherence [31].
2. A coherence-driven agent architecture for norm-autonomous agents together with an algorithm for coherence-driven agent reasoning.

The remainder of the paper is structured as follows. We first give a general introduction to Thagard's 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 a generic coherence framework that can be used to create coherence-driven agents. We discuss in this framework how pieces of information can be organised in the form of a graph, along with the necessary computable functions to evaluate and maximise the coherence of such a graph. We then specialise the formulation for a particular type of coherence, namely deductive coherence. We derive a deductive coherence function based on the deduction relation of a logic, although the function we define is independent of the underlying logic. Next, we introduce a proof-theoretic characterisation of coherence focusing on deductive coherence. We discuss the formal properties of coherence and illustrate how these properties help us to derive coherence values between pieces of information.

Then, we define a coherence-driven agent as a cognitive agent whose decisions or actions are based on coherence maximisation. For this purpose we define certain specific graphs corresponding to a cognitive agent. We adapt concepts from multi-context systems so that our coherence-driven agent can reason with its cognitions and norms put together. We later sketch a procedure an agent may follow in the context of a normative multiagent system.

Finally we conduct a case study, in which we take a specific normative multiagent system. This case study is 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. One of the regions modelled as a coherence-driven agent decides to adopt or violate the norm (in this case, the signed treaty) driven by its coherence maximisation.

2 Theory of Coherence

In this section, we introduce the theory of coherence in general and provide a summary of Thagard's theory, which is the major inspiration and the base of this work. We then interpret Thagard's theory as a decision theory and contrast it with other decision theories.

2.1 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 important when studying the structure of how knowledge is acquired or beliefs are 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 foundationalist rejects answers 1 and 2 and argues 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 [17]. Coherentism, however, argues that the second 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 that 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 admit 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 do not need justification). This sometimes is described as the meeting point between foundationalism and coherentism [24].

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 [27]. 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 paper bases its foundations on this theory, we introduce it next.

2.1.1 Thagard's Formalisation

Paul Thagard is one of the philosophers who have attempted to introduce a computational interpretation of coherence. Thagard postulates that the theory of coherence is a cognitive theory with foundations in philosophy that approaches problems in terms of the satisfaction of multiple constraints within networks of highly interconnected elements [31, 32]. At the interpretation level, Thagard's theory of coherence is the study of associations, that is, how a piece of information influences another and how best different pieces of information can fit together. Each piece of information imposes constraints on others, the constraints being positive or negative. Positive constraints strengthen pieces of information, thereby increasing coherence, while negative constraints weaken them, thereby increasing incoherence. Hence, a coherence problem is to put together those pieces of information that have a positive constraint between them, while separating those having a negative constraint. Coherence is maximised if we obtain such a partition of information where a maximum number of constraints is satisfied. Thus in this sense, reasoning is framed as a classification problem.

Thagard formalises coherence as follows [31]: The basic notions are that of a set of pieces of information that are represented as nodes in a graph with weighted links, or constraints, between these nodes. Further, some of these constraints are positive (representing coherence) and others negative (representing incoherence), and associated with each constraint is a number that indicates the weight of the constraint. Given these, maximising coherence is formulated as the problem of partitioning the set of nodes into two sets, $\mathcal{A}$ (the accepted nodes) and $\mathcal{R}$ (the rejected nodes), in a way that it maximises compliance with the following two coherence conditions:

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

If an edge complies with one of the above conditions, then, Thagard defines it as a satisfied constraint. The coherence problem is thus simply to maximise the sum of the weights of the satisfied constraints.

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

Thagard has also experimented with many computational implementations of coherence. ECHO is a computational model of explanatory coherence which uses a connectionist algorithm [31]. Though there is no guarantee that such neural network models for coherence would converge to a coherence-maximising partition, he claims that on small networks it has been shown to give good results.

Thus, Thagard proposes the first major concrete account of coherence that takes us from the abstract notion of coherence to a computational phenomenon that can be evaluated. One of the main drawbacks 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 paper is mainly an attempt in this direction, while also attempting to propose the theory as a primary decision making mechanism for agents in normative multiagent systems.

2.2 Comparison with Other Decision Theories

Keeping Thagard's approach to coherence as maximising constraint satisfaction, we try to understand the main concept behind this theory. We associate coherence with an ever-changing system where coherence is the only property that is preserved, while everything around it changes. In cognitive terms, this would mean that, there are no beliefs nor other 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. However, we do not suppose that taking decisions based on coherence imply an unstable system. Our claim is based on the fact that some beliefs are more fundamental than others. Revision of such fundamental 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 while the process of coherence maximisation further helps resolve conflicts by selecting among the best alternatives.

When applied to decision making, this means that we may not only select the set of actions to be performed to achieve certain fixed goals, but also look for the best set of goals to be pursued. Further, since coherence affects everything from beliefs to goals and actions, it may happen that beliefs contradicting a decision made are discarded. There are psychological theories such as cognitive dissonance [18] that explain this phenomenon as an attempt to justify the action chosen. Thus, with coherence we are looking at a more dynamic model of cognitions where one picks and chooses goals, actions and even beliefs to fit a grand plan of maximising coherence. In concrete terms, a highly desired state of the world (desired in a classical sense) may get discarded in front of a less desired state of the world because it is incoherent with the rest of the beliefs, desires or intentions.

As discussed in [31], this view of decision making is very different from those of classical decision-making theories where the notion of *preference* is atomic and there is no conceptual understanding of how preferences can be formed. In contrast, coherence-based decision making tries to understand and evaluate these preferences from the available complex network of constraints. The assumption here is more basic because the only knowledge available to us are the various interacting constraints between pieces of information.

3 Coherence Framework

In this section we introduce the generic coherence framework together with those computable functions that will allow us to build coherence-driven 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 [9, 27] as we modify 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 cognitions of an agent. In the following definitions, we introduce what we call coherence graphs, the various computable functions to determine the coherence of such graphs, and how the coherence of a given graph can be maximised.

3.1 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 propositions representing concepts, actions or mental states both atomic and complex, graded and absolute. Edges between nodes may be associated with a strength, represented by a function ζ, 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 ζ assigns a positive strength to the edge connecting those pieces of information. Based on Thagard's classification of the types of coherence, we have different ζ functions. The value of the function ζ, 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. Thagard's principles may be used to define a function ζ for each different type of coherence (see Section 4).

We consider a running example as in Figure 1, which will help us to illustrate the concepts as we define them. The graph in the example captures deductive coherence as yielded by classical propositional deduction. As we gradually build our framework, we shall see how these coherence values arise.

Definition 3.1
A *coherence graph* is an edge-weighted undirected graph $g = \langle V, E, \zeta \rangle$, where

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

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

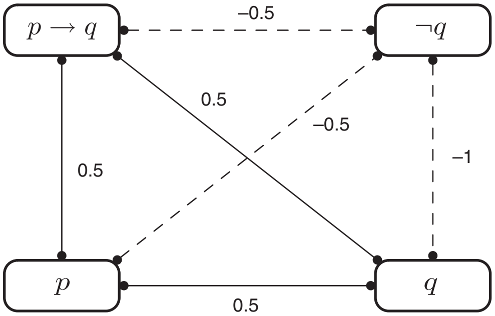

Fig. 1. An example of a coherence graph

3.2 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 3.2
Given a coherence graph $g = \langle V, E, \zeta \rangle$, and a partition $(\mathcal{A}, V \setminus \mathcal{A})$ of V, the *set of satisfied constraints* $C_{\mathcal{A}} \subseteq E$ is given by

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

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

To illustrate this, consider the partition as in Figure 2. We see that $\{p \rightarrow q, \neg q\}$, $\{p \rightarrow q, p\}$ and $\{p, \neg q\}$ are the satisfied constraints. Now we define the coherence-maximising partition as the partition that maximises the satisfaction of constraints. We first define the strength of a partition as the sum over the strengths (the ζ values) of all the satisfied constraints divided by the the number of edges in the graph. Then the coherence-maximising partition is that which maximises the strength, and the coherence value of the graph is defined as the strength of this partition.

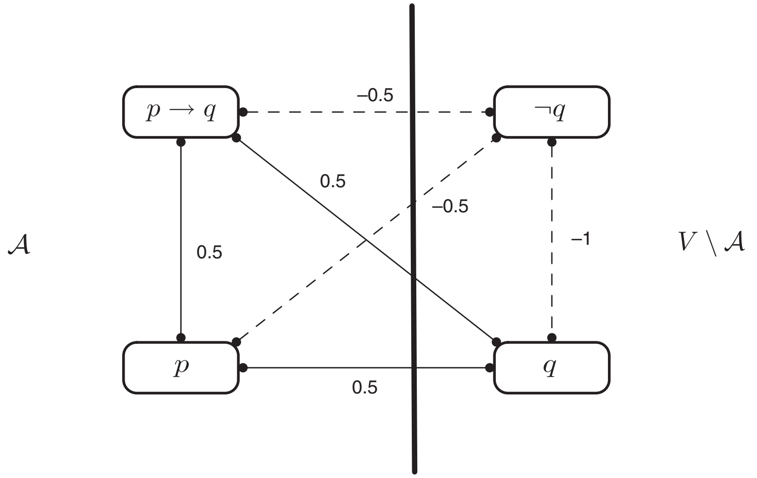

Fig. 2. A partition of a coherence graph

Definition 3.3
Given a coherence graph $g = \langle V, E, \zeta \rangle$, *the strength of a partition* $(\mathcal{A}, V \setminus \mathcal{A})$ of V is given by

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

For the partition in Figure 2, the strength is 0.25. Notice that, by Definitions 3.2 and 3.3,

$$\sigma(g, \mathcal{A}) = \sigma(g, V \setminus \mathcal{A}) \ . \tag{1}$$

Definition 3.4
Given a coherence graph $g = \langle V, E, \zeta \rangle$, *the coherence of* g is given by

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

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

Due to Equation 3.1, the accepted set $\mathcal{A}$ is never unique for a coherence graph. Moreover, there could be other partitions that generate the same value for $\kappa(g)$. Here we mention some possible criteria for selecting an accepted set among the alternatives. If $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ are sets from all those partitions that maximise coherence of the graph g, first we may choose the accepted set to which the intuitively obvious propositions belong. This is based on one of Thagard's principles (which we will formalise in the next definition) on deductive coherence [31], namely that *intuitively obvious propositions have an acceptability on their own*. Further, the coherences of the sub-graphs $g|_{\mathcal{A}_i}$, $i \in [1, n]$ can give us an indication of how strongly connected they are. The higher the coherence, the more preferred the corresponding accepted set may be. And lastly, an accepted set with more number of elements could be preferred to another with less.

The coherence maximising partition for our example is as in Figure 3. With this partition we see that all constraints are satisfied, and this partition gives the maximum strength for the graph, namely $0.58\bar{3}$.

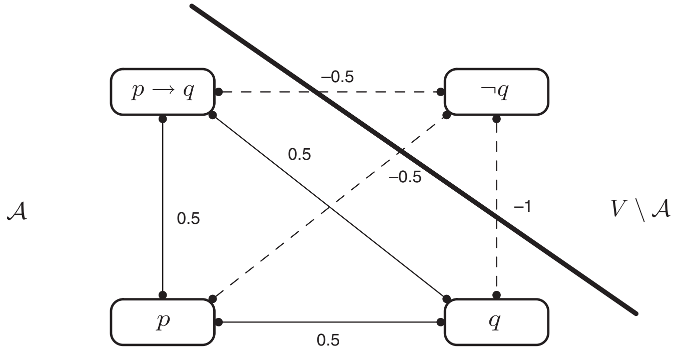

Fig. 3. Coherence-maximising partition of a coherence graph

4 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 possibly some associated confidence degrees), we need to define a coherence function ζ that assigns a coherence value to pairs of pieces of information. As the nature of relationship between two pieces of information can vary greatly, we may need to define more than one coherence function. But for each type of coherence, only one such relationship is evaluated. That is, for explanatory coherence, two pieces of information are coherent only if they are related by an explanation.

Here we study one such type of coherence, namely deductive coherence, and define a *deductive coherence function* ζ which captures the deductive relationship between propositions. We choose deductive coherence among the different types of coherence because logical deduction has a sound theoretical basis and has well defined rules in order to start with a formalisation of coherence. We first derive a deductive coherence function in adherence with Thagard's principles and 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.

We base our coherence function on multiset deduction 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 such as linear logics, relevance logics or multi-valued logics. We shall abbreviate "multiset deduction relation" with MDR. We assume that all MDRs we deal with are finitary and decidable. These MDRs are often called *simple consequence relations* [5]. We define an MDR as follows:

Definition 4.1
Given a logical language L, a *multiset deduction relation (MDR)* on L, is a binary relation $\vdash$ between finite multisets of formulas of L such that, for all $\Gamma, \Gamma_1, \Gamma_2, \Sigma_1, \Sigma_2 \subseteq L$ and for all $\gamma \in L$:

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

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

4.1 Coherence Functions

Thagard introduces in [31] 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.

Before going into the details of Thagard's principles, it is important to note that these principles were proposed taking into account a context or —in logical terminology— a theory $\mathcal{T}$. Examples of such theories may be the theory of arithmetic while proving theorems in mathematics, or legal laws while making legal judgements. In the context of autonomous normative agents, the set of rules and observations about a context is this theory. To be rigorous we should call $\mathcal{T}$ a finite theory presentation. However, to avoid lengthy phrases, we will often call it just a theory. Assuming bounded rationality for our agents, $\mathcal{T}$ is not closed under deduction. We essentially see the process of coherence maximisation as a process of theory revision. That is, each time the agent encounters a new piece of information β (a new norm, a new belief, etc.), it tries to relate it to the theory presentation it has. The new information can influence $\mathcal{T}$ in the following ways:

1. *Extend $\mathcal{T}$: β helps to deduce propositions that were not deducible before.*
2. *Extend $\mathcal{T}$: β is deducible from $\mathcal{T}$.*
3. *Modify $\mathcal{T}$: β is in a deduction relation with some propositions in $\mathcal{T}$, however, contradicts some other.*

The coherence function we propose here is in the context of a theory $\mathcal{T}$ and is motivated to aid this process of theory revision. We use Thagard's principles to relate an MDR with a coherence function ζ. Principles 2 and 3 capture the fact that there are certain positive coherence relations between premises, and between each of the premises and the conclusion. Since we want to focus only on those deductions that fall in the context of the theory, we shall define the coherence only between those formulas that are either in the theory or are subformulas thereof (hence Definition 4.2 below). Principle 4 gives an indication of the magnitude of coherence. It states that it decreases with the increasing number of premises required. Principle 5 discusses the case of contradiction. And finally, Principle 7 stresses the basic notion of coherence, namely that if anything is accepted, it is because accepting it improves the coherence of the system. Therefore, the theory $\mathcal{T}$ is also part of our coherence graph, and its acceptance is only with respect to coherence maximisation.

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

Definition 4.2
Let L be a logical language and let $\mathcal{T} \subseteq L$ be a finite theory presentation. We call $\mathcal{T}^\bullet$ the *closure of $\mathcal{T}$ under subformulas* when $\alpha' \in \mathcal{T}^\bullet$ if and only if there is an $\alpha \in \mathcal{T}$ such that α' is a subformula of α.

Definition 4.3
Let L be a logical language and $\vdash$ be an MDR for L. Let $\mathcal{T} \subseteq L$ be a finite theory presentation. A *support function* $\eta_{\mathcal{T}}: \mathcal{T}^\bullet \times \mathcal{T}^\bullet \to [-1,1] \setminus \{0\}$ with respect to $\mathcal{T}$ is given by:

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

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

Definition 4.4
Let L be a logical language and $\vdash$ be an MDR for L. Let $\mathcal{T} \subseteq L$ be a finite theory presentation. Let $\eta_{\mathcal{T}}: \mathcal{T}^\bullet \times \mathcal{T}^\bullet \to [-1,1] \setminus \{0\}$ be a support function with respect to $\mathcal{T}$. A *deductive coherence function* $\zeta_{\mathcal{T}}: (\mathcal{T}^\bullet)^{(2)} \to [-1,1] \setminus \{0\}$ with respect to $\mathcal{T}$ is given by:

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

Our example of Figure 1 assumes that $\mathcal{T} = \{p \to q, \neg q\}$, and consequently $\mathcal{T}^\bullet = \{p \to q, \neg q, p, q\}$. The only relevant deductions using formulas of $\mathcal{T}^\bullet$ (and assuming classical propositional deduction) are:

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

Therefore, we have that

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

and consequently,

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

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

PROPOSITION 4.5
The deductive coherence function $\zeta_{\mathcal{T}}$ as defined in Definition 4.4 satisfies Thagard's principles of deductive coherence (see Section 4.1).

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

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

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

Let $\Gamma_1,\Gamma_2\subseteq\mathcal{T}^\bullet$ with $|\Gamma_1|<|\Gamma_2|$, and let $\alpha_1,\alpha_2,\beta\in\mathcal{T}^\bullet$ such that $\Gamma_1,\alpha_1,\vdash\beta$ and there does not exists Γ'_1 such that $|\Gamma'_1|<|\Gamma_1|$ and $\Gamma'_1,\alpha_1,\vdash\beta$, and $\Gamma_2,\alpha_2\vdash\beta$ and there does not exists Γ'_2 such that $|\Gamma'_2|<|\Gamma_2|$ and $\Gamma'_2,\alpha_2,\vdash\beta$, and $\Gamma_1\nvdash\beta$ and $\Gamma_2\nvdash\beta$ and $\alpha_1\nvdash$ and $\alpha_2\nvdash$. Then $\eta_{\mathcal{T}}(\alpha_1,\beta)>\eta_{\mathcal{T}}(\alpha_2,\beta)$, and consequently, $\zeta_{\mathcal{T}}(\{\alpha_1,\beta\})>\zeta_{\mathcal{T}}(\{\alpha_2,\beta\})$, which satisfies Principle 4.

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

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

Finally, Definition 3.4 satisfies Principle 7. ∎

4.2 Properties of Coherence Based on MDRs

To conclude this section, we explore the properties of the deductive coherence function $\zeta_{\mathcal{T}}$ that would help us determine the values for pairs of formulas related through some of the structural rules and connectives of the underlying logic. We do this by identifying the properties of the support function $\eta_{\mathcal{T}}$ using the properties of the connectives and structural rules. Our aim is to stress the generality of our approach, although these properties are not essential for the understanding of the coherence-driven agent architecture introduced in Section 5.

We can classify logics according to structural rules (such as weakening or monotonicity) and connectives available in it. There are two types of connectives: the *internal* connectives, which transform a given sequent into an equivalent one that has a special required form, and the *combining* connectives, which combine two sequents into one. For instance, classical propositional logic is monotonic, satisfies weakening, has all internal and combining connectives, and makes no difference between them. On the other hand, propositional linear logic is monotonic, has also all connectives, but distinguishes between internal and combining ones. Intuitionistic logic differs from classical propositional logic in its implication connective and does not have internal negation.

By Definition 4.3, the function $\eta_{\mathcal{T}}$ is defined for formulas of $\mathcal{T}^{\bullet} \subseteq L$ related through an MDR in the form $\Gamma, \alpha \vdash \beta$. Hence we express the deduction relation in this single-conclusioned form so that we can find properties of function $\eta_{\mathcal{T}}$ between different formulas of the premises and conclusion, using the properties of the connectives.[2]

4.2.1 Combining Conjunction

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

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

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

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

4.2.2 Internal Conjunction

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

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

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

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

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

4.2.3 Combining Disjunction

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

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

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

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

4.2.4 Internal Disjunction

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

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

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^\bullet$ and $\alpha + \beta, \gamma \in \mathcal{T}^\bullet$ such that $\Gamma \nvdash \alpha$ and $\Gamma \nvdash \alpha + \beta$ and $\alpha \nvdash$ and $\beta \nvdash$ and $\gamma \nvdash$. Further, let $\vdash$ satisfy Weakening.[3]

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

4.2.5 Combining Implication

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

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

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

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^\bullet$ and $\alpha \supset \beta, \sigma \in \mathcal{T}^\bullet$ such that $\Gamma \not\vdash \sigma$ and $\alpha \supset \beta \not\vdash$ and $\beta \not\vdash$.

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

4.2.6 Internal Implication

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

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

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^\bullet$ and $\alpha \rightarrow \beta, \gamma \in \mathcal{T}^\bullet$ such that $\Gamma, \gamma \not\vdash \beta$ and $\Gamma \not\vdash \alpha \rightarrow \beta$ and $\alpha \not\vdash$ and $\gamma \not\vdash$.

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

4.2.7 Internal Negation

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

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

Let $\mathcal{T} \subseteq L$ and $\Gamma \subseteq \mathcal{T}^\bullet$ and $\neg\alpha, \gamma \in \mathcal{T}^\bullet$ such that $\Gamma, \alpha \not\vdash$ and $\Gamma, \gamma \not\vdash$.

1. if $\eta(\gamma, \alpha) < 0$ then $\Gamma, \gamma, \alpha \vdash$, and consequently $\Gamma, \gamma \vdash \neg\alpha$ and hence $\eta(\gamma, \neg\alpha) \geq -\eta(\gamma, \alpha)$.
2. If $\Gamma, \gamma \vdash \neg\alpha$ then $\eta(\gamma, \neg\alpha) > 0$, and since $\Gamma, \gamma, \alpha \vdash$ we have that $\eta(\gamma, \alpha) \leq -\eta(\gamma, \neg\alpha)$.
3. $\eta(\neg\alpha, \alpha) = -1$

5 An Architecture for Coherence-Driven Agents

In this section we describe an architecture for coherence-driven agents based on the coherence framework developed so far. A *coherence-driven agent* is an agent that always takes an action based on maximisation of coherence of its cognitions, norms, and other social commitments. We further consider cognitive agents such as those based on BDI theory [29], since it is one of the prominent existing agent architectures. We use an adaptation of the architecture developed by Casali et al. [11], which is based on multi-context systems (MCS) and incorporates reasoning under uncertainty.

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 between theories. Contexts in a multi-context BDI are the contexts of belief, desire, and intention cognitions. The deduction mechanism of MCS is based on two kinds of inference rules, internal rule inside each context, and bridge rules between contexts. Internal rules allow an agent to draw consequences within a context, and they determine an MDR, while bridge rules allow to embed results from one context into another [20, 21]. Thus, we shall define an agent as a tuple $\langle \{C_i\}_{i=1\ldots n}, B\rangle$ consisting of:

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

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

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

In our adaptation of the multi-context architecture, the theories $\mathcal{T}_i$ of the contexts will yield coherence graphs. We have already defined the coherence function ζ derived from an MDR $\vdash_i$ within one context (see Definition 4.4). In the following we define how this coherence function is to be extended to capture how coherence arises between formulas due to bridge rules carrying consequences from one graph to another. For this we will define two additional kinds of functions, graph-extension and graph-join functions. First we begin by giving a brief overview of the contexts in our multi-context system before defining these functions.

5.1 Cognitive and Norm Contexts

Here we discuss the belief, desire, intention, and norm contexts corresponding to a normative BDI agent. We take the belief, desire, and intention contexts as defined in Casali et al. [10]. For the norm logic associated with the norm context, we use the work of Godo et al. [15] on probabilistic deontic logic. We here give a sketch of a belief context, while the details are in Casali et al. [11]. The desire, intention, and norm contexts can be defined in a similar fashion, with the belief logic replaced either by a desire, intention, or norm logic, accordingly. Further, the belief theory $\mathcal{T}_B$ gives rise to a coherence graph whose nodes are graded formulas of the belief language.

5.1.1 Belief Logic

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

If φ is a proposition in L, the intended meaning of $B\varphi$ is that "φ is believable". We use a modal many-valued logic based on Łukasiewicz logic to formalise $\mathcal{K}_B$ as follows: [4]

1. Given a propositional language L, the belief language L_B of $\mathcal{K}_B$ is given by:
   - If $\varphi \in L$ then $B\varphi \in L_B$
   - If $r \in \mathbb{Q} \cap [0,1]$ then $\bar{r} \in L_B$
   - If $\Phi, \Psi \in L_B$ then $\Phi \rightarrow_L \Psi \in L_B$ and $\Phi \,\&\, \Psi \in L_B$ (where & and $\rightarrow_L$ correspond to conjunction and implication of Łukasiewicz logic)

   Other Łukasiewicz logic connectives for the modal formulas can be defined from &, $\rightarrow_L$ and $\bar{0}$: $\neg_L \Phi$ is defined as $\Phi \rightarrow_L \bar{0}$, $\Phi \wedge \Psi$ as $\Phi \,\&\, (\Phi \rightarrow_L \Psi)$, $\Phi \vee \Psi$ as $\neg_L(\neg_L \Phi \wedge \neg_L \Psi)$, and $\Phi \equiv \Psi$ as $(\Phi \rightarrow_L \Psi) \,\&\, (\Psi \rightarrow_L \Phi)$.
2. The axioms A_B of $\mathcal{K}_B$ are:
   - all axioms of propositional logic;
   - the axioms of Łukasiewicz logic for modal formulas (for instance, axioms of Hájek's Basic Logic (BL) [22] plus the axiom $\neg_{BL} \neg_{BL} \Phi \rightarrow_{BL} \Phi$);
   - the probabilistic axioms, i.e., given $\varphi, \psi \in L$,

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

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

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

5.1.2 Belief Graph

Our aim is, given a theory presentation $\mathcal{T}_B \subseteq L_B^*$ expressed using graded beliefs in a belief language, to define the corresponding coherence graph (see Definition 5.4 further below). For this, however, we first need to extend the definitions given in Section 4 for graded beliefs underling a belief language L_B. The idea is to determine coherence values between graded beliefs not only by virtue of the deduction relation; we will also take into account the grades as specified in the theory presentation $\mathcal{T}_B \subseteq L_B^*$.

Definition 5.1
Let L_B be the belief language as defined above, and let $\mathcal{T}_B \subseteq L_B^*$ be a finite theory presentation using only graded formulas. We call $\mathcal{T}_B^\bullet$ the *closure of* $\mathcal{T}_B$ *under subformulas* when $(B\varphi', r') \in \mathcal{T}_B^\bullet$ if and only if there is an $(B\varphi, r) \in \mathcal{T}_B$ such that φ' is a subformula of φ and $r' = \sup\{q \mid \mathcal{T}_B \models (B\varphi', q)\}$.

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

Definition 5.2
Let L_B be the belief language and $\vdash_B$ the MDR as defined above. Let $\mathcal{T}_B \subseteq L_B^*$ be a finite theory presentation using only graded formulas. A *support function* $\eta_{\mathcal{T}_B} : \mathcal{T}_B^{\bullet} \times \mathcal{T}_B^{\bullet} \rightarrow [-1,1]$ with respect to $\mathcal{T}_B$ is given by:

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

Definition 5.3
Let L_B be the belief language as defined above. Let $\mathcal{T}_B \subseteq L_B^*$ be a finite theory presentation using only graded formulas. Let $\eta_{\mathcal{T}_B} : \mathcal{T}_B^{\bullet} \times \mathcal{T}_B^{\bullet} \rightarrow [-1,1] \setminus \{0\}$ be a support function with respect to $\mathcal{T}_B$. A *deductive coherence function* $\zeta_{\mathcal{T}_B} : (\mathcal{T}_B^{\bullet})^{(2)} \rightarrow [-1,1] \setminus \{0\}$ with respect to $\mathcal{T}_B$ is given by:

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

Definition 5.4
Let $\mathcal{K}_B = \langle L_B, A_B, \vdash_B \rangle$ be a belief logic, where L_B is a belief language, A_B are a set of axioms and $\vdash_B$ is an MDR. Let $\mathcal{T}_B \subseteq L_B^*$ be a finite theory presentation expressed using graded beliefs only. A *belief graph* of $\mathcal{T}_B$ is the coherence graph $g = \langle V, E, \zeta \rangle$ where

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

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

5.1.3 Norm Graph

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 to understand the relative importance of a norm within a system of norms. A graded norm is interpreted in terms of its priority, measured in terms of the value it generates in a normative multiagent system. This value can be determined by the social goals it helps in achieving. However, there could be other measures for determining the 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 Probability-valued Deontic Logic [15], to represent and reason with norms. We define the norm language L_N by extending a classical propositional language L defined upon a countable set of propositional variables and connectives $\neg$ and $\rightarrow$. In particular, L_N is defined as a fuzzy modal language over Standard Deontic Logic (SDL) to reason about the probability degree of deontic propositions. In our case the probability values are replaced by grades associated with norms. The language, axioms and deduction relation are defined similarly as in the case of the belief logic, and again we will be interested in the sublanguage L_N^* of graded formulas. For details, we refer to [15].

Some examples of formulas in a graded normative language L_N^* are given below. Also, to keep uniformity with the belief, desire, and intention languages as described above, we adopt a slightly different notation from that given in [15]:

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

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

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

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

5.2 Reasoning Across Contexts

Reasoning in a BDI normative agent architecture needs to consider the influence of cognitions and norms between each other. For instance, it is desirable to choose or predict an action that is most coherent with the set of cognitions and accepted norms. It is also desirable to know the influence of a new information on the overall coherence of cognitions and norms. Typically, in a multi-context system, such reasoning is achieved by the use of bridge rules. For coherence-driven agents we adapt the idea of bridge rules to be able to establish links and coherence values across several coherence graphs.

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

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

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

Definition 5.5
We say that a function $\varepsilon:\mathcal{G}^n\rightarrow\mathcal{G}^n$, $n>0$, is a *graph-extension function* if, given a tuple of graphs $\bar{g}=\langle g_1,\dots,g_n\rangle$ in $\mathcal{G}^n$, $\varepsilon(\bar{g})=\bar{g}'$ is such that

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

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

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

Definition 5.6
We say that a sequence is an *extension sequence* if, given a tuple of graphs $\bar{g}\in\mathcal{G}^n$, $n>0$, and a set of graph-extension functions $S\subseteq\mathcal{E}$,

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

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

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

Definition 5.7
We say that a function $\iota:\mathcal{G}^n\rightarrow\mathcal{G}$, $n>0$, is a *graph-join function* if, given a tuple of graphs $\bar{g}=\langle g_1,\dots,g_n\rangle$ in $\mathcal{G}^n$, with $g_i=\langle V_i,E_i,\zeta_i\rangle$, $\iota(\bar{g})=\langle V,E,\zeta\rangle$ is such that:

- $V=\bigcup_{1\leq i\leq n}\{i:\Phi\mid\Phi\in V_i\}$

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

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

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

Definition 5.8
We say that a function $\varsigma:\mathcal{G}^n\to\mathcal{G}$, $n>0$, is a *graph-composition function* if, given a tuple of graphs $\bar{g}$ in $\mathcal{G}^n$, a set of graph-extension functions $S\subseteq\mathcal{E}$ with a unique fixed point and a graph-join function $\iota\in\mathcal{J}$, $\varsigma(\bar{g}) = \iota(S^*(\bar{g}))$.

5.2.1 Bridge Rules — Composition Functions

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

Definition 5.9
Given a family $\{C_i\}_{i=1\ldots n}$ of contexts, $n>0$, a bridge rule b is a rule of the form

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

where:

- i_k (with $k\in\{1,\ldots,q\}$ and $q>0$) and j are all pairwise distinct[5] indeces of contexts (i.e., $1\leq i_k, j\leq n$)
- A_k and A are are formula schemata specifying premises from contexts C_{i_k} and a conclusion from context C_j, respectively
- R_k are either variables or numerical constants in $\mathbb{Q}\cap[0,1]$
- $f(R_1,R_2,\ldots,R_q)$ is an expression, where $f:(\mathbb{Q}\cap[0,1])^q\to\mathbb{Q}\cap[0,1]$

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

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

Definition 5.10
Let $\{C_i\}_{i=1\ldots n}$ be a family of contexts, $n>0$, and let

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

[5] This condition can be dispensed with; we only require it for subsequent ease of presentation of Definition 5.11 below.

be a a bridge rule as in Definition 5.9. The graph-extension function ε_b is defined as follows: Given a tuple of graphs $\bar{g} = \langle g_1, \ldots, g_j, \ldots, g_n \rangle$ (with $g_j = \langle V_j, E_j, \zeta_j \rangle$), then $\varepsilon_b(\bar{g}) = \langle g_1, \ldots, g_j', \ldots, g_n \rangle$ where $g_j' = \langle V_j', E_j', \zeta_j' \rangle$ such that

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

Given a finite set of bridge rules $B \subseteq \mathcal{B}$, we derive a graph-join function that, given tuple $\bar{g} = \langle g_1, \ldots, g_n \rangle$ of graphs, joins all graphs g_i together, adding new edges and their coherence values between nodes corresponding to instances of premise schemata A_k and the conclusion schema A, and also between instances of premise schemata themselves, in accordance to Thagard's principles discussed in Section 4.1.

Definition 5.11
Let $\{C_i\}_{i=1\ldots n}$ be a family of contexts, $n > 0$, and let $B \subseteq \mathcal{B}$ be a finite subset of bridge rules. The graph-join function ι_B is defined as follows: Given a tuple of graphs $\bar{g} = \langle g_1, \ldots, g_n \rangle$ (with $g_i = \langle V_i, E_i, \zeta_i \rangle$), then $\iota_B(\bar{g}) = \langle V, E, \zeta \rangle$ such that

- $V = \bigcup_{1 \le i \le n} \{ i : \Phi \mid \Phi \in V_i \}$
- $$E = \bigcup_{1 \le i \le n} \Big\{ \{ i : \Phi, i : \Psi \} \mid \{\Phi, \Psi\} \in E_i \Big\} \cup$$
$$\bigcup_{b \in B} \left\{ \{ i : \Phi, j : \Psi \} \;\middle|\; \begin{array}{l} i : \Phi \text{ is a premise of } \pi(b) \text{ and } j : \Psi \text{ is the conclusion of } \pi(b), \\ \text{where } \pi \text{ is a most general substitution, such that,} \\ \text{for all premises } k : (A, R) \text{ of } b,\ \pi((A, R)) \in V_k \end{array} \right\} \cup$$
$$\bigcup_{b \in B} \left\{ \{ i : \Phi, j : \Psi \} \;\middle|\; \begin{array}{l} i : \Phi \text{ and } j : \Psi \text{ are premises of } \pi(b),\ i \neq j, \\ \text{where } \pi \text{ is a most general substitution, such that,} \\ \text{for all premises } k : (A, R) \text{ of } b,\ \pi((A, R)) \in V_k \end{array} \right\}$$
- $\zeta(\{ i : \Phi, i : \Psi \}) = \zeta_i(\{\Phi, \Psi\})$, and $\zeta(\{ i : \Phi, j : \Psi \})$ for $i \neq j$ is defined as in Definition 5.3 with respect to the following support function:

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

5.2.2 Application of Composition Functions — An Example

Consider, for example, the tuple of graphs $\langle g_B, g_D, g_I\rangle$ corresponding to a multi-context system with belief context C_B, desire context C_D and intention context C_I, and with a single bridge rule

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

Let's further assume that $(Bp, 0.95)$ is a node in g_B and $(Dp, 0.7)$ is a node in g_D where p is a proposition. Then,

- $\varepsilon_b(\langle g_B, g_D, g_I\rangle) = \langle g_B, g_D, g'_I\rangle$, where g'_I is g_I with the node $(Ip, 0.7)$ added to its set of nodes.
- $\iota_{\{b\}}(\langle g_B, g_D, g'_I\rangle)$ is the disjoint union of graphs g_B, g_D, and g'_I with the additional edges $\{B\!:\!(Bp, 0.95), I\!:\!(Ip, 0.7)\}$, $\{D\!:\!(Dp, 0.7), I\!:\!(Ip, 0.7)\}$, and $\{B\!:\!(Bp, 0.95), D\!:\!(Dp, 0.7)\}$, with coherence value 0.35 for all of them.

5.3 Coherence-Driven Agents

Equipped with contexts and a mechanism to reason across these contexts, we can now 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 $C_i = \langle L_i, A_i, \vdash_i, \mathcal{T}_i\rangle$ where L_i, A_i and $\vdash_i$ are the language, axioms, and inference rules of a logical system, and $\mathcal{T}_i$ is a finite theory presentation. In our extension of MCS, a coherence-driven agent will further have a function $\mathsf{cohgraph}$ that maps a context to its corresponding coherence graphs. And a function $\mathsf{compfun}$ that maps a set of bridge rules to a graph-composition function. This extension is required because contexts are expressed as coherence graphs and agents will need both coherence and graph-composition functions to reason within and between graphs. For the normative BDI agents considered here, the contexts are C_B, C_D, C_I and C_N, 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 5.12
A *coherence-driven agent* a is a tuple $\langle\{C_i\}_{i=B,D,I,N}, B, \mathsf{cohgraph}, \mathsf{compfun}\rangle$ where $\{C_i\}_{i=B,D,I,N}$ is a family of contexts, $B \subseteq \mathcal{B}$ is a set of bridge rules, $\mathsf{cohgraph}\!:\!\{C_i\}_{i=B,D,I,N} \to \mathcal{G}$ maps contexts to coherence graphs, and $\mathsf{compfun}\!:\!2^B \to \mathcal{G}^{(\mathcal{G}^4)}$ maps sets of bridge rules to composition functions that take a quadruple of graphs to a graph.

In the following we describe how coherence-driven agents interact with a normative environment. As in Figure 4, a coherence-driven agent starts with a set $\{C_i\}_{i=B,D,I,N}$ of contexts corresponding to the beliefs, desires, intentions and norms, which it expresses as coherence graphs. We assume that the languages of each context are all extensions of the same propositional language L. It is also desirable for the theory presentations of contexts to be consistent. Our proposal, however, is tolerant to inconsistencies and in a certain sense exposes them and eliminates them, if possible. Further, the agent is assumed to have a set B of bridge rules to reason across contexts and it computes the composite graph using these.

An agent at any time can either perceive the normative environment or make a prediction about a future action. In the event of a new piece of information $(K\varphi, r)$ (where K is one of

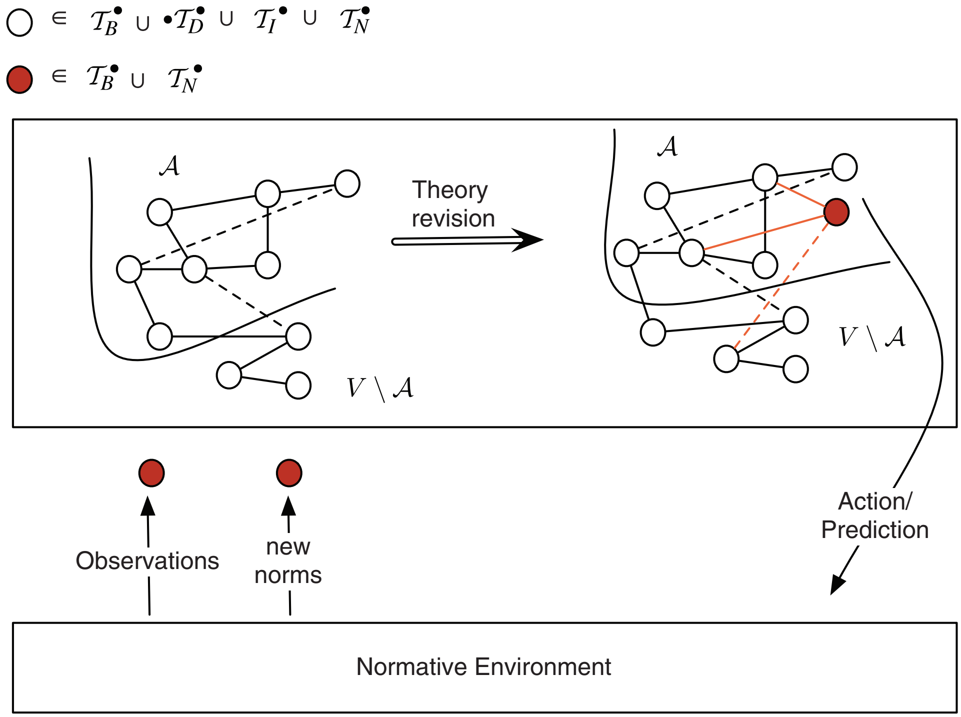
○ ∈ $\mathcal{T}_B^\bullet \cup {}^\bullet\mathcal{T}_D^\bullet \cup \mathcal{T}_I^\bullet \cup \mathcal{T}_N^\bullet$
● ∈ $\mathcal{T}_B^\bullet \cup \mathcal{T}_N^\bullet$
$\mathcal{A}$
Theory revision
$\mathcal{A}$
$V \setminus \mathcal{A}$
$V \setminus \mathcal{A}$
Observations
new norms
Action/ Prediction
Normative Environment

FIG. 4. Architecture of a coherence-driven agent

the modal operator of its context languages) the agent reasoning proceeds according to the following algorithm:

1. it adds the new graded formula to the theory $\mathcal{T}$ of the corresponding context C_B, C_D, C_I, or C_N (depending if K is either B, D, I, or else O or P, respectively);
2. it computes the deductive closure of $\mathcal{T}$ (but, to keep the closure finite, we compute it limited to $\mathcal{T}^\bullet$, i.e., without introducing formulas with new subformulas that are not in $\mathcal{T}$);
3. it expresses the contexts with newly closed theories as coherence graphs computing the tuple $\bar{g} = \langle \mathsf{cohgraph}(C_B), \mathsf{cohgraph}(C_D), \mathsf{cohgraph}(C_I), \mathsf{cohgraph}(C_N)\rangle$ associated to them;
4. it computes the composite graph $g = \mathsf{compfun}(\bar{g}) = \iota_B(S^*(\bar{g}))$, where $S = \{\varepsilon_b \mid b \in B\}$;
5. it computes all coherence-maximising partitions[6] $(\mathcal{A}, V \setminus \mathcal{A})$, where V is the set of nodes of g, by computing $\operatorname{argmax}_{\mathcal{A} \subseteq V} \sigma(g, \mathcal{A})$, and eventually chooses one of them;[7]
6. it updates the theory presentations of the contexts according to the newly accepted set.

As discussed in Section 4, if the new piece of information $(K\varphi, r)$ reinforces the original theory, it is added to the accepted set and the theory becomes more coherent. If it contradicts

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

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

elements of the original theory, then either the new piece of information is rejected, or some part of the already accepted theory is rejected, whichever makes the theory more coherent. To make predictions, however, the agent uses only the accepted theory. This is realistic, as it is the accepted set that the agent wants to base its decisions on.

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

6 Example — Water Sharing Conflict

We apply the formalism developed so far in a real context of political conflict. We show how a cognitive agent endowed with a coherence management system is capable of taking decisions by means of maximising the coherence of its beliefs, desires and intentions. The example is motivated by the water sharing treaty signed between the southern states of India during 1892 and 1924 and the disputes thereafter [1]. The objectives of this example are twofold. First, to demonstrate how self-interested agents evaluate norms in a BDI context. Second, to show a process of *norm adoption* caused by changes in the individual coherence evaluations as a result of new information (beliefs, desires, etc.) being acquired. The grand aim of this case study is to set up a framework for *norm adaptation* (i.e., norm change) itself, which will be part of our future work.

We describe now the reasoning performed by a coherence-driven agent in such a setting. We simplify the case for brevity: we model the reasoning of just one of the agents (southern Indian state s) involved in the conflict in three snapshots of time 1891, 1892, and 1991, the first one when the first treaty is about to be signed (when the decision to adopt a norm is to be taken), the second, when the norm is adopted and the third after a long period of co-operation between the states, when the situation had significantly evolved and the norm is to be broken by s.

6.1 Terminology

To represent the cognitions and norms of an agent, we shall use belief, desire, intention and norm languages as defined in Section 5.1. Hence, $(B\varphi, r)$ represents that the agent believes that proposition φ is true with degree at least r. (Propositions $(D\varphi, r)$, and $(I\varphi, r)$ are desires and intentions and are interpreted analogously.) $(O\varphi, r)$ is the obligation of the agent to make φ happen. The degree r is a measure on the relevance of the norm, such as for instance its priority, or to what extend it needs to be fulfilled. The statements about the world are in propositional language where each proposition is a grounded predicate with obvious meaning as can be seen later in the snapshots.

We assume that the agent has four contexts C_B, C_D,C_I, and C_N containing the beliefs, desires, intentions, and permissions/obligations, respectively. We will omit the reference to the contexts in the notation as all beliefs are in C_B, desires in C_D, intentions in C_I, and permissions and obligations in C_N, and therefore there is no possible confusion. The two bridge rules we use in the water-sharing example are the following:

- $b_1 = \frac{(B\varphi,r) \quad (D\varphi,s)}{(I\varphi,\min(r,s))}$: Whenever a proposition is believed with degree at least r and desired with degree at least s, then a corresponding intention with a degree at least $\min(r,s)$ is added to the theory of context $C_I I$. We don't intend stronger than we desire or we believe.
- $b_2 = \frac{(B\varphi,r) \quad (O\varphi,s)}{(I\varphi,\min(r,s))}$: If the agent beliefs that an obligation is feasible, then it intends to make it happen.

6.2 Pre-treaty situation (1891)

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

In Table 1, we list the elements of the theories $\mathcal{T}_B, \mathcal{T}_D$ and the deductive closure of context C_B for agent s before the norm was proposed. The propositions φ_i are as given in Table 2. The coherence graph, g_1 obtained from these theories is represented in Figure 5 at the end of Step 5 of the agent reasoning process (see Section 5.3), that is, including the effects

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

TABLE 1. s's theories and the deductive closure of context B. In this case only the belief context deduces new formulas.

φ_1	$good(rainfall)$
φ_2	$adequate(waterlevel)$
φ_3	$satisfied(demand)$
φ_4	$\varphi_1 \wedge \varphi_2$
φ_5	$\varphi_4 \rightarrow \varphi_3$

TABLE 2. Propositions relevant for s's cognitions at the beginning of the reasoning.

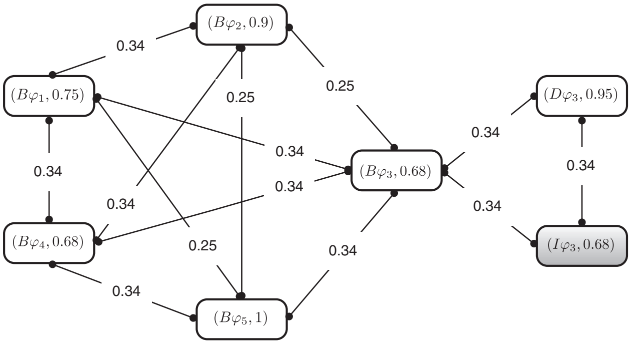

FIG. 5. Initial coherence graph (g_1) of s as in 1891 including the bridge rule deductions (shadowed) with $\kappa(g_1) = 0.32$

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

Table 3. New elements introduced into s's theories in 1892

φ_6	$\mathit{release}(\mathit{300\ billion\ ft^3})$
φ_7	$\mathit{realised}(\mathit{attack})$
φ_8	$\varphi_1 \wedge \varphi_2 \wedge \varphi_6$
φ_9	$\varphi_8 \rightarrow \neg\varphi_3$
φ_{10}	$\neg\varphi_6 \rightarrow \varphi_7$

Table 4. Propositions relevant for s's cognitions in 1892

of the deductions by means of bridge rules (i.e. $(I\varphi_3,0.68)$). The coherence $\kappa(g_1)$ is 0.32 and the accepted set $\mathcal{A}$ includes all the nodes in the graph. The computation is illustrated in one case: $\eta((B\varphi_1,0.75),(B\varphi_2,0.9)) = max(\frac{0.65}{2}) = 0.38$, given that $(B\varphi_1,0.75),(B\varphi_2,0.9) \vdash (B\varphi_4,0.75*0.9)$ (assuming probabilistic independence) where Γ is empty, and there is no other possible deduction to obtain one of the formulas from the other.

6.2.1 Norm adoption. Evaluating the Treaty (1892).

In 1892, a new norm was proposed: the Indian state s will get obliged to release 300 billion ft^3 of water to its neighbour state annually; included in the proposal there was a threaten of military retaliation in the case of unfulfillment of the obligation. Certainly, the release of water might threaten the objective of satisfying the internal demand and state s was not necessarily happy with it. The situation of the theories at the beginning of the treaty is as expressed in Table 3. We have incorporated the deduced intention of the previous subsection as there was no conflict with any previous intention (the intention theory was empty —Step 6 in the algorithm). Also, we have added the new formulas associated with the obligation and its related facts.

Agent s evaluates the proposal of the new treaty by incorporating into its theories and its respective coherence graphs the new obligation, its implications and the sanctions that might be incurred if the proposal is not accepted. That is, the theories are updated according to Table 3, where the relevant propositions φ_i are as in Table 4.

Agent s now computes the composite coherence graph g_2 (shown in Figure 6) resulting from the theory update and using the set of bridge rules $B = \{b_1, b_2\}$. There are no negative coherence values between any pair of cognitions so the whole set is accepted again. However, this time the overall strength of the maximal partition is $\kappa(g)$ is 0.225. It is clear that coherence has decreased by incorporating the new norm which might be interpreted as an indication that the overall situation for s was not as good as before signing the treaty. But still the accepted set includes the acceptance of the norm. Hence, guided by coherence maximisation, agent s signs the treaty.

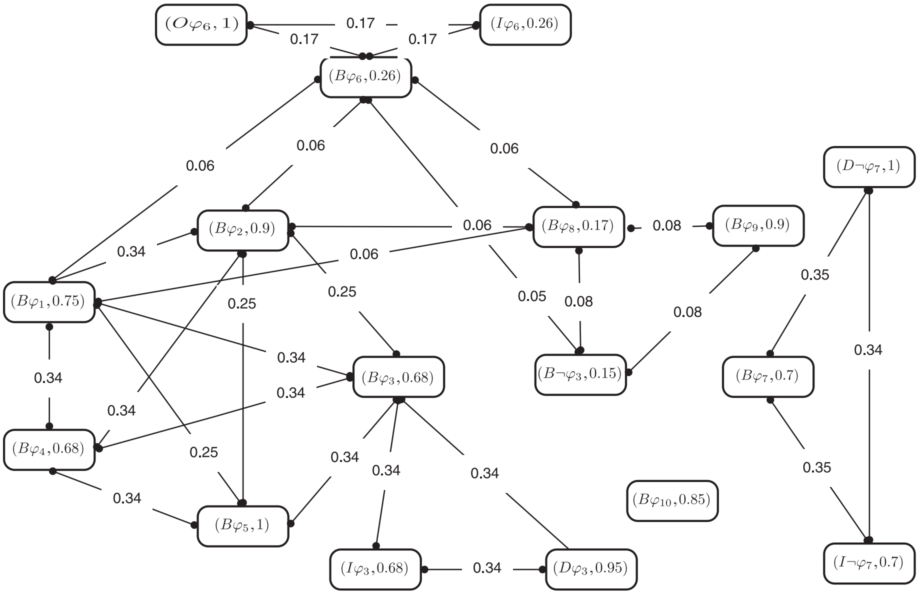

FIG. 6. Coherence graph (g_2), with norm accepted $\kappa(g_1)=0.225$

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

TABLE 5. New elements introduced into s's theories in 1991

6.3 The Incoherence Buildup (1991)

By 1991 s experiences large-scale industrialisation, urbanisation, and higher revenue growth and as a consequence s also experiences higher water usage. Specially important for the example is the fact that an increase in water usage means that the possibility of satisfying the internal demand will decrease as the fact $(B(\varphi_{11}\to\neg\varphi_3),0.8)$ indicates (see Table 5).

In this last time point we don't depict the complete graph (see Figure 7) as it is a bit large and we just show some of the relevant nodes where the incoherence has built up. The facts involved are listed in Table 6.

The incoherence builds up as the degree of belief in $\neg\varphi_3$ increases as a consequence of the new added facts.[8] That creates a negative coherence that forces for the fist time a partition with maximum coherence that does not include all the nodes in the graph. Using the criterion of selecting as the accepted set of the partition the subgraph with maximal strength the agent chooses the set on the right as the accepted one. In particular, node

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

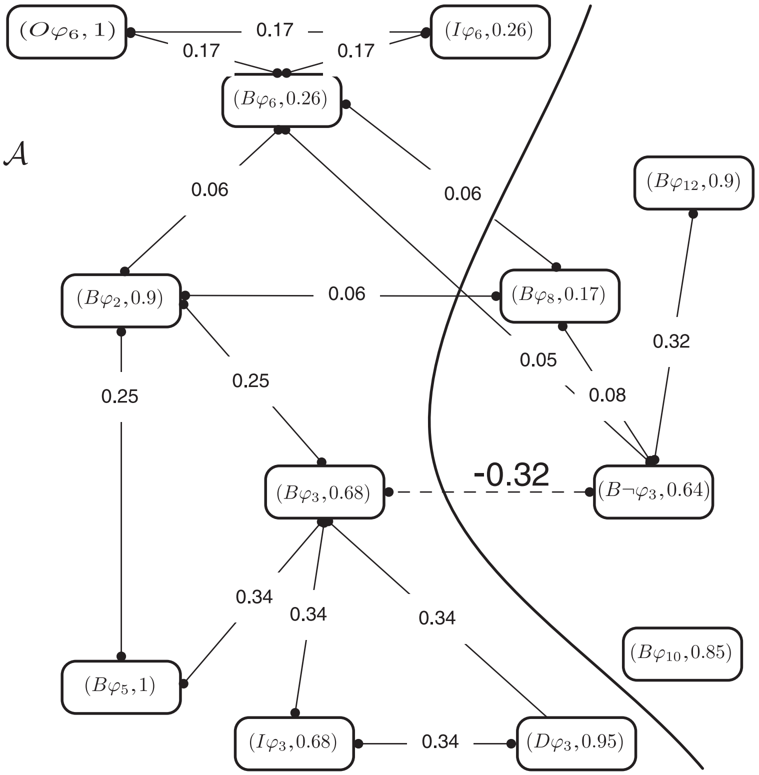

FIG. 7. Subgraph of the coherence graph (g_3)

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

TABLE 6. Propositions relevant for s_1's cognitions in 1991

$(O\varphi_6, 1)$ appears in the rejected set and therefore the Indian state s decides to break the norm in order to keep a maximal coherence value.

6.4 Discussion

Even though the example only demonstrates the case of a single norm, the example can be easily extended to cases where there are multiple norms and there is a need to choose among them. In terms of coherence, this is again selecting a norm (or a set of norms) that maximises the coherence of the graph. By inserting all the norms into the coherence graph and then calculating the coherence maximising partition, we can see which of the norms fall into the accepted set and hence can be adopted together. Another point to note is that here we have assumed our agents to be coherence maximising. But in reality there may be other criteria that need to be considered. Some of them we already mentioned are sanctions and rewards. Another important factor by which an agent makes a decision to adopt a norm is observing the behaviour of other agents. We can represent this by adding nodes into the graph that represent our beliefs on the cognitive stance of other agents.

6.5 Computational Complexity

It has been shown that it is possible to convert a coherence maximisation problem into an equivalent max-cut problem [31]. 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 approximations. Thagard in his formalisation of coherence gave several implementations of coherence, with an extensive implementation of a neural network model called ECHO [31]. He also compared it with a max-cut implementation. We have extended Thagard's implementation to incorporate additional features of our model. That is, our implementation uses 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 of Definition 5.2. 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 [34]. Experimental evaluation of the case study is part of our immediate future work.

7 State of the Art

In this paper, we proposed a coherence-based framework to design increasingly sophisticated agents with autonomous capabilities, and illustrated that such agents would take flexible and dynamic decisions when faced with dynamic and uncertain scenarios. We particularly aimed at introducing autonomous agents in the context of normative multiagent systems and demonstrated, from the point of view of an agent, that autonomy helps in evaluating norms (rather than following designer specifications without deliberation). On the other hand, we attempted to formalise the theory of coherence in a generic and computationally plausible manner and, to a large extend, independent of the domain of interest. Because our work links different areas of research, we explore in this section the work done in few of those important related areas, namely autonomous agent deliberation, normative systems and autonomous norm evaluation, and formalisation of coherence. Further, argumentation has been a popular means for internal and external deliberation in agents and hence treated as an important means to bring in autonomy in agents. Hence we make a comparison of our coherence-based framework with argumentation frameworks and remark about a few interesting works in the field of legal reasoning where argumentation has been the predominant mechanism for decision making.

7.1 Autonomous Agent Deliberation

From the years that agent theory came into existence, autonomy is one of the most desired features to be incorporated in agent design. The first major step was made when a behaviour model of agents was proposed —the BDI model for artificial agents based on the theory of rational action in humans put forward in 1988 by the philosopher M. Bratman [8]. The BDI model 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 [29]. 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) [35] to BOID [9], 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 of agents true autonomous entities.

The work of Pasquier et al. [27] is an attempt at bringing more autonomous and dynamic 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 model of cognitive coherence that 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 [31]. 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 developing a coherence framework that is fully computable and generic, and in particular, on exploring methods to compute coherence values between pieces of information.

7.2 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 [7]. One of the early introductions of norms for multiagent co-ordination is the work on artificial social systems by Tennenholtz et al. [19, 25, 30]. 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 [12, 14, 36]. 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 Boella et al. [6] 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 Boella 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 [13, 14]. 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. 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 that can be embedded in the process of norm emergence proposed by the authors.

7.3 Formalising Coherence

Here we primarily analyse those proposals that formalise coherence. The theory of coherence has been studied in philosophy, computer science, and law. There are very few attempts to formalise coherence so that it could be used as a general framework. Still, there exist 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 stay the same. We compare and contrast their proposals and our work.

The work of Piwek attempts to model dialogue coherence in terms of generative systems based on natural deduction [28]. The main argument in his work is that it is possible to generate coherent dialogues by relying on entailment relations in the agent's knowledge base. The paper primarily deals with information-seeking dialogues where the definition of whether an agent knows a fact is equated to whether this fact 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 modelled. 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. models agent dialogues based on the theory of dissonance [33]. 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 have not considered the fact that not all incoherences are dissonances [27, 33]. 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.

7.4 Comparison with Argumentation Frameworks

Since the work of Pollock, Loui, and others, argumentation systems are a popular means to study non-monotonic reasoning. Apart from studying non-monotonicity, argumentation systems are also increasingly used to model deliberation, negotiation and decision making. For example, the frameworks of Bondarenko, Dung, Toni and Kowalski are used to model agent deliberation both from internal and from external perspectives [4, 26, 27]. Internal deliberation helps agents to deal with internal conflicts in goals, and conflicts among goals and norms. External deliberation assists a group of agents to reach consensus or agree on decisions through persuasion and negotiation. Since Dung's framework is the most abstract framework studied and widely used in particular argumentation systems, we highlight its main characteristics and contrast it with our work.

An argument system is characterised by pairwise attack relations between arguments. The concept of acceptability and admissibility of arguments are central notions in the theory. Acceptability of an argument with respect to a set is a measure of its justification within the set, however, note that it is a Boolean measure. Admissible sets are those that have all elements as accepted with respect to that set. Further, the notion of a preferred extension finds the maximal over the admissible sets. Stable extensions are preferred extensions with yet another constraint, namely that every argument outside of it is defeated. These notions, in a broad sense, capture the idea of a maximal, conflict-free, and justified set of arguments. In coherence terms, a preferred or stable extension can be compared to accepted sets of a coherence graph.

However, there are a few important differences that distinguish a preferred set of arguments from that of an accepted set of a coherence graph. First of all, a preferred extension attempts to find those sets of arguments that are indisputable. They do not tolerate inconsistencies or contradictions. An argument based system tends to be more brittle in that it cannot easily cope with varying degrees of acceptability. Usually it is an all or nothing affair: given a set of arguments, an argument is either acceptable or not; there is nothing in between. Whereas in reality, an argumentation system may need to find the least inconsistent set of arguments than to find the absolute set that is justified. Coherence maximisation finds such a set with tolerance to inconsistencies.

Another difference is that often arguments contradict one another not in absolute values. One can associate a degree of contradiction, or a degree of support between arguments. It is important to account for this degree as they have an impact on deciding the final outcome. Since coherence captures this degree of relatedness between elements, coherence may give us a more realistic measure.

However, argument-based approaches yield explicit reasons why an outcome should be adopted and explicit refutations of alternatives. Coherence-based approaches are criticised for their lack of transparency. However, in our approach, since we derive our coherence measures from the deduction relation of an underlying logic, we make explicit the process, why two pieces of information are related (or why two arguments are on an attack relation). Further, since coherence maximisation is interpreted as maximising satisfaction of constraints, we pick a set as an accepted set when it satisfies maximum constraints, or when we have a maximal set of arguments that are in some sense justified and coherent. This process has the added advantage that it not only looks for justified arguments but coherent arguments.

In [3], Amaya tries to apply a notion of coherence in legal justification and studies how notions of fairness and coherence are related. The work also claims that coherence considerations need to be taken into account while putting forward an argument along with truth and fairness considerations. In her work, Amaya analyses Thagard's model of coherence as constraint satisfaction and argues that such a model should 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 a formalisation. Another work on argumentation applies a coherence-based mechanism for practical reasoning systems [16].

From the above discussion on the related work, it is clear that there is both a need for psychologically inspired models such as cognitive theory of coherence and an interest in developing such systems from diverse areas. The need is however for formalisation of the abstract theories and for developing computationally feasible models.

8 Conclusion and Future Work

In this paper, we have proposed a coherence-based framework which extends the popular BDI architecture by including the notion of coherence. The first research objective we set out to achieve in this paper was (see Section 1):

> "to find mechanisms for reasoning in norm-autonomous agents that enable them to resolve conflicts among cognitions and norms."

We proposed coherence maximisation as a reasoning mechanism for adopting or violating norms. To build coherence maximising agents, we proposed a coherence framework that incorporates the representation, and coherence maximising functions to enable actual computation. We also defined computable functions to generate coherence constraints in the coherence graph of an agent. We further provided an algorithm that lets an agent take decisions regarding norm adoption. We illustrated this with the help of an example drawn from a real world scenario.

The second research objective we set out to achieve in this paper was (see Section 1):

> "to determine if and how such a reasoning tool can be incorporated in an agent theory such as BDI theory."

We defined an agent architecture that extends a BDI-based one by introducing coherence maximisation as the principle reasoning mechanism. In particular, we extended a multi-context BDI architecture to incorporate the fact that theories in each context are expressed as coherence graphs. We also proposed an algorithm to implement coherence-driven agents that accept or rejects a new piece of information driven by coherence maximisation. There are, however, many associations and details to be explored further. We mention some of them in the discussion below.

In the paper we mention that we see the process of coherence maximisation as the process of theory revision. The intuition is that, if a belief in the accepted set moves to the rejected set, the confidence on the belief no longer remains the same and should be reduced. The contrary is true if a rejected belief is accepted again. Thus, the degrees on the cognitions or norms can be computed using a probabilistic distribution driven by coherence maximisation. However, this topic needs to be further explored and its connection with other popular theories such as AGM [2] needs to be studied.

The choice of deductive coherence is another point of discussion. We admit that for agent reasoning and decision making, probably an analysis of deliberative or explanatory coherence would be much more appropriate. We picked deductive coherence so that we could readily show the significance of incorporating coherence in an agent theory. In this case, the underlying deduction relation is well understood. However, we could apply our results to different types of coherence.

There are also criticisms raised about the philosophy of coherence. One question often put forward with respect to the application of coherence as an agent decision-making mechanism 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. One of our main future work is to place coherence among other theories of rationality and justify the above claims with concrete results.

Another important question is about the computational feasibility of coherence maximisation. Unlike other proposals on coherence maximisation, in this paper, we have introduced a fully computational framework of coherence, although coherence maximisation is an NP-complete problem. However, as we assume bounded rationality for our agents, our coherence graphs are bounded and are not complete. To further reduce the computational burden, our future work aims to bring in contexts that would consider only sub-graphs of the actual ones, with the intuition that coherence maximisation should consider only those nodes that are relevant to the problem at hand.

An important issue we have not explored in depth is the treatment of norms. We would like to study the structure of norms in more detail, and possibly express them as coherence constraints on the cognitions. This would enable agents in a normative system to generate new norms, and further help agents to evaluate existing norms in the context of its cognitions.

Finally, in the present work we have dealt with the cognitive aspect of a normative multiagent system. In the future we would like to explore norm evolution in a society of coherence-driven agents. In particular, we plan to study how agents can agree upon a set of norms, and adapt them when required, and explore equilibrium conditions for coherence.

Acknowledgements

The authors wish to express their thanks to Francesc Esteva, for his comments and ideas on the semantic approach to coherence, and the reviewers of this paper for their helpful comments. The research presented in this article was partially supported by OpenKnowledge

(STREP FP6-027253) funded by the European Commission; by Agreement Technologies (CONSOLIDER-INGENIO 2010 CSD2007-0022) and MULOG2 (TIN2007-68005-C04-01) funded by Spain's *Ministerio de Ciencia e Innovación*; and by the *Generalitat de Catalunya* (2005-SGR-00093).

References

[1] Kaveri river water dispute. In *Wikipedia, The Free Encyclopedia*, October 27, 2008. Retrieved December 8, 2008, from http://en.wikipedia.org/wiki/Kaveri_River_Water_Dispute.

[2] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.

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

[4] Leila Amgoud, Nicolas Maudet, and Simon Parsons. Modeling dialogues using argumentation. In *4th International Conference on Multi-Agent Systems (ICMAS 2000), 10-12 July 2000, Boston, MA, USA*, pages 31–38. IEEE Computer Society, 2000.

[5] Arnon Avron. Simple consequence relations. *Information and Computation*, 92(1), 1991.

[6] 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 2004)*, pages 483–486, 2004.

[7] Guido Boella, Leendert van der Torre, and Harko Verhagen. Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, 12(2–3):71–79, 2006.

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

[9] Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the 5th International Conference on Autonomous Agents*, pages 9–16. ACM, 2001.

[10] Ana Casali, Llus Godo, and Carles Sierra. Graded BDI models for agent architectures. In *Computational Logic in Multi-Agent Systems, 5th International Workshop, CLIMA V, Lisbon, Portugal, September 29-30, 2004, Revised Selected and Invited Papers*, *Lecture Notes in Computer Science*, vol. 3487, pages 126–143. Springer, 2005.

[11] 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 Computación*, 2006.

[12] Cristiano Castelfranchi, Frank Dignum, Catholijn M. Jonker, and Jan Treur. Deliberative normative agents: Principles and architecture. In *Intelligent Agents VI, Agent Theories, Architectures, and Languages (ATAL), 6th International Workshop, ATAL '99, Orlando, Florida, USA, July 15-17, 1999, Proceedings*, *Lecture Notes in Computer Science*, vol. 1757, pages 364–378. Springer, 2000.

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

[14] Rosaria Conte, Cristiano Castelfranchi, and Frank Dignum. Autonomous norm acceptance. In *Intelligent Agents V, Agent Theories, Architectures, and Languages, 5th International Workshop, ATAL '98, Paris, France, July 4-7, 1998, Proceedings*, *Lecture Notes in Computer Science*, vol. 1555, Springer, 1999.

[15] Pilar Dellunde and Lluis Godo. Introducing grades in deontic logics. In *Deontic Logic in Computer Science, 9th International Conference, DEON 2008, Luxembourg, Luxembourg, July 15-18, 2008. Proceedings*, *Lecture Notes in Computer Science*, vol. 5076, pages 248–262. Springer, 2008.

[16] Paul E. Dunne and Trevor J. M. Bench-Capon. Coherence in finite argument systems. *Artificial Intelligence*, 141(1):187–203, 2002.

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

[18] Leon Festinger. *A Theory of Cognitive Dissonance.* Stanford University Press, 1957.

[19] David Fitoussi and Moshe Tennenholtz. Choosing social laws for multi-agent systems: Minimality and simplicity. *Artificial Intelligence*, 119(1-2):61–101, 2000.

[20] Fausto Giunchiglia. Contextual reasoning. *Epistemologia (Special Issue on I Linguaggi e le Macchine)*, XVI:345–364, 1993.

[21] Fausto Giunchiglia and Luciano Serafini. Multilanguage hierarchical logics or: How we can do without modal logics. *Artificial Intelligence*, 65(1):29–70, 1994.

[22] Petr Hájek. *Metamathematics of Fuzzy Logic*, *Trends in Logic*, vol. 4, Kluwer Academic Publishers, 1998.

[23] Martin J. Kollingbaum and Timothy J. Norman. Norm adoption in the NoA agent architecture. In *The Second International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2003, July 14-18, 2003, Melbourne, Victoria, Australia, Proceedings*, pages 1038–1039. ACM, 2003.

[24] Paul K. Moser(ed.). *The Oxford Handbook of Epistemology.* Oxford University Press, 2002.

[25] Yoram Moses and Moshe Tennenholtz. Artificial social systems. *Computers and Artificial Intelligence*, 14(6):533–562, 1995.

[26] Simon Parsons, Carles Sierra, and Nick R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8:261–292, 1998.

[27] Philippe Pasquier and Brahim Chaib-draa. The cognitive coherence approach for agent communication pragmatics. In *The Second International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2003, July 14-18, 2003, Melbourne, Victoria, Australia, Proceedings*, pages 544–551. ACM, 2003.

[28] Paul Piwek. Meaning and dialogue coherence: A proof-theoretic investigation. *Journal of Logic, Language and Information*, 16(4):403–421, 2007.

[29] Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multiagent Systems, June 12-14, 1995, San Francisco, California, USA*, pages 312–319. MIT Press, 1995.

[30] Yoav Shoham and Moshe Tennenholtz. On social laws for artificial agent societies: Offline design. *Artificial Intelligence*, 73(1-2):231–252, 1995.

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

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

[33] Erika Valencia and Jean-Paul Sansonnet. Model for dialogue between informational agents. In *Progress in Artificial Intelligence, 11th Protuguese Conference on Artificial Intelligence, EPIA 2003, Beja, Portugal, December 4-7, 2003, Proceedings*, *Lecture Notes in Computer Science*, vol. 2902, pages 355–359. Springer, 2003.

[34] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

[35] Michael Wooldridge. *Reasoning about Rational Agents*. MIT Press, 2000.

[36] Fabiola López y López, Michael Luck, and Mark d'Inverno. Constraining autonomy through norms. In *The First International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2002, July 15-19, 2002, Bologna, Italy, Proceedings*, pages 674–681. ACM, 2002.

Received 8 December 2008