Formalising Deductive Coherence: An Application to Norm Evaluation

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

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

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

PILAR DELLUNDE, *Univ. Autònoma de Barcelona 08193 Bellaterra, 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. Cognitive coherence in particular is the coherence theory based explanation of the mind which evaluates the truth of a cognition in terms of it being a member of some suitably defined body of other cognitions: a body that is consistent, coherent, and possibly endowed with other virtues, provided these are not defined in terms of truth. Thus a coherent set is interdependent such that every element in it contributes to the coherence. We take the Thagard's proposal of a coherence set as that of maximizing satisfaction of constraints between elements and explore its use in normative multiagent systems. In particular we show how it could be used as a mechanism to introduce true autonomy in agents particularly in a normative multi agent setting. We show, how an agent can deliberate on norm evaluation while also considering other factors such as sanctions and rewards.

We first provide a general coherence framework with the necessary computing tools. Later we introduce a proof-theoretic characterization 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, thus proposing a fully computational model of coherence. Our use of graded logic helps us to incorporate reasoning under uncertainty which is more realistic in the context of multiagent systems. We then explore a scenario where agents deliberate about norm evaluation 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 these systems, norms are identified with obligations, permissions and prohibitions, which specify the ideal behaviour of agents. Such systems also consider constitutive norms to give a new meaning to certain behaviour of agents. While a normative multiagent system is prescriptive about agent behaviour, it does so within the framework of autonomous agents. That is, the system assumes agents to behave in an autonomous manner and reason about norms autonomously. This is because of the fact that the success of such systems does not depend on all the agents following the prescribed norms blindly, rather on having rational agents deliberating about the prescribed norms, evaluating their usefulness, selectively following those norms that improve their efficiency, and effecting a change when there are conflicts, inefficiencies, or situational changes that they can perceive.

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

We are certainly not the first to identify this need, and there have been numerous attempts in the recent past explicitly addressing this issue [22, 12, 4, 24, 34, 20]. Many of these efforts are focused towards extending the cognitive agent theory (for instance the Belief, Desire, and Intention (BDI) theory) with explicit representation of norms such as in BOID [7], EMIL [12], and NoA [20], or propose a more comprehensive multiagent system architecture that is norm-aware as in [4]. However, apart from providing static-priority based autonomy[1] and recognising autonomous norm acceptance phases, a gap still exists to enrich agent theories with true autonomy.

To enhance the autonomous capabilities of agents, we propose a normative agent theory which extends the BDI theory with the theory of coherence [29]. Coherence theory, when used to explain human reasoning, proposes that humans accept or reject a cognition (external or internal) depending on how much it contributes to maximise the constraints imposed by situations and other cognitions. Pasquier et al. [24] introduced the possibility of extending agent reasoning with Thagard's theory of coherence. While their contribution introduces the concept of coherence in the field of multiagent systems, they still do not clarify the nature of a coherence relation nor do they specify how a coherence graph can be constructed. Thus, a general treatment of coherence to be used to realise computational models is still called for.

According to the theory of coherence, there are coherence and incoherence relations between *pieces of information* depending on whether they support (yielding a positive constraint) or contradict (yielding a negative constraint) each other. If two pieces of information are not related, then, there is no coherence (constraint) between them. Due to the fact that coherence is evaluated based on constraints that exist between

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

pairs of information, a graph representation is most intuitive. Normally a graph with nodes and weighted edges are used to represent the pieces of information and constraints between them. Given such a coherence graph, Thagard defines a mechanism to compute the overall coherence of the graph based on maximising constraint satisfaction between pairs of nodes. Certain principles are also defined to characterise and differentiate various types of coherence relations that might exist between pairs. Understanding these principles and deducing methods to compute the coherence values between them is fundamental to compute the overall coherence of a given coherence graph. Without this important formalisation, practical realisations of coherence are hard to imagine.

In this 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. Our aim is to generate coherence values between pairs of information (in this case, formulas in a logical language) by formalising the relationship between coherence and logical entailment. Coherence as a logical relation is significant in itself and has important implications: it is tolerant to inconsistencies and allows us to work with deductive systems without certain structural rules such as weakening.

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

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

We address the first question by following other researchers [24] to propose the theory of coherence for autonomous normative agent design. Though the theory has been proposed earlier to extend BDI agents in the context of communication, it has not been proposed as a general theory in the context of normative systems. We address the second question by proposing a clear method, illustrating how a coherence-driven agent can reason autonomously about norms and cognitions by the process of coherence maximisation. We list the important steps in the process as follows.
Given a set of graded agent cognitions and a set of graded norms (norms with priorities) of a normative multiagent system,

1. evaluate the coherence or incoherence relations between pairs of cognitions of the same type and between pairs of norms;
2. evaluate coherence or incoherence relations between pairs of cognitions of differing types and norms;
3. given that all possible coherence or incoherence relations are computed for cognitions and norms, evaluate the overall coherence of the agent as if the agent were to accept all the cognitions and all the norms;
4. separate the set of cognitions and norms into two sets by a process of coherence maximisation such that only elements of one set are considered to be accepted (considered valid);

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

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 which can be used to create coherence-driven agents. We discuss in this framework, how pieces of information can be organised in the form of a graph, along with the necessary computable functions to evaluate and maximise the coherence of such a graph. We then specialise the formulation for a particular type of coherence, namely deductive coherence. We derive a deductive coherence function based on the deduction relation of a logic, however the function we define is independent of the underlying logic. Later 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.

Later, we define a coherence-driven agent as a cognitive agent whose utility maximisation is achieved by coherence maximisation. For this purpose we define certain specific graphs corresponding to a cognitive agent. We adapt concepts from multi-context systems so that 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. We show that by coherence maximisation how the agent evaluate norms (in this case, the signed treaty) which leads to adopt or violate norms.

2 Theory of Coherence

In this section, we introduce the theory of coherence and provide a summary of Thagard's Theory of Coherence, which is the major inspiration and the base of this 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 an important problem when studying the structure of how knowledge is acquired or belief is justified. One of the central questions in the regress problem is to know how one knows or is justified in believing some particular thing. Many epistemologists studying justification have

attempted to argue for various types of chains of reasoning that can escape the regress problem.

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

There are two main schools of thought in answering this question, foundationalism and coherentism. The foundationlist reject answers 1 and 2 and argue that 3 is the valid answer. According to the foundationalist option, the series of beliefs terminates with special justified beliefs called basic beliefs: these beliefs do not owe their justification to any other beliefs from which they are inferred [15]. 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 which makes it impossible for two entirely different sets of beliefs to be internally coherent. Thus, there might be several such sets, and pure coherentism does not offer a solution. However later theories of coherence admits certain favorable statements whose presence in a set makes it more coherent than other competing sets. These special statements are some of the obvious statements (which does not need justification). This sometimes is described as the meeting point between foundationalism and coherentism [21].

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

2.2 Thagard's Theory of Coherence

Thagard postulates that coherence theory is a cognitive theory with foundations in philosophy that approaches problems in terms of the satisfaction of multiple constraints within networks of highly interconnected elements [29, 30]. Thagard takes the theory from its abstract form and gives concrete interpretations of it. More importantly Thagard attempts to extend the theories reach to a broad audience by explaining how the theory of constraint satisfaction can be applied to problems of probabilistic reasoning, social consensus, emotions, and decision making in general. Though his argument about coherence being a theory of everything is not fully con-

vincing, he makes a strong case for specific uses of the theory in concrete problems of decision making and probabilistic reasoning.

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

2.2.1 Thagard's Formalisation

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

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

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

By itself, this characterisation has no philosophical or psychological or probabilistic reasoning applications, because it does not state the nature of the elements, the nature of the constraints, or the algorithms to be used to maximise satisfaction of the constraints.However, Thagard further proposes that there are six main kinds of coherence: explanatory, deductive, conceptual, analogical, perceptual, and deliberative, each with its own array of elements and constraints. Once these elements and constraints are specified, then the algorithms that solve the general coherence problem can be used to compute coherence in ways that apply specific domain problems.

2.2.2 Computing Coherence

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

complexity and effectiveness. They are as given below:

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

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

Thus, Thagard proposes the first major concrete account of coherence, which takes us from the abstract notion of coherence to a a computational phenomena which can be evaluated. One of the main drawback of his theory is that, he stops with giving certain principles about calculating values of coherence constraints for different types of coherence. However, to compute these values, one needs to have concrete functions with proven properties. This paper is mainly an attempt in this direction, while also attempting to propose the theory as a primary decision mechanism for agents in normative multiagent systems.

2.3 Comparison with Other Decision Theories

Keeping Thagard's approach to coherence as maximising constraint satisfaction, we try to understand the main concept behind this theory. We associate coherence with an ever-changing system where coherence is the only property that is preserved, while everything around it changes. That is, everything else is picked and chosen to maximise coherence. In cognitive terms, this would mean that, there are no beliefs nor other cognitions that are taken for granted or fixed forever. Everything can be changed and may be changed to keep coherence. We humans tend to revise or re-evaluate adherence to social norms, our plans, goals and even beliefs when we are faced with incoherence. For most researchers of agent theory and multiagent systems, however, changing beliefs in this way is equivalent to creating agents that are not dependable. However we argue that taking decisions based on coherence does not imply an unstable system. Our claim is based on the fact that some beliefs are more fundamental than others. Such fundamental beliefs define a personality, and revision of a fundamental belief is less frequent compared to other beliefs. In coherence terms, these beliefs are fundamental because they support and get support from most other cognitions and hence are in positive coherence with them. Hence, such beliefs will almost always be part of the chosen set while maximising coherence. The same is the case with other cognitions. Those that are normally in positive coherence with most other, will almost always be selected with coherence maximisation, while this process also helps us resolve conflicts by selecting the best alternatives.

When applied to decision making, this means, we not only select the set of actions to be performed to achieve certain fixed goals, but we also look for the best set of goals

to be pursued. Further, since coherence affects everything from beliefs to goals and actions, it may happen that beliefs contradicting a decision made are discarded. There are psychological theories such as cognitive dissonance that explain this phenomenon as an attempt to justify the action chosen. Thus, with coherence we are looking at a more dynamic model of cognitions where one picks and choses goals, actions and even beliefs to fit a grand plan of maximising coherence.

As discussed in [29], this view of decision making is very different from those of classical decision making theories. The fundamental notion of classical decision theories is the notion of *preference*. The notion of *utility* is derived from the notion of preference in such a way that x is preferred to y if and only if x has a greater utility than y. Then, the decision making process is equated to the maximisation of utilities. However, preferences are atomic, and there is no conceptual understanding of how preferences are formed. In the theory of coherence, we precisely aim at understanding preferences. The assumption here is more basic because the only knowledge available to us are the various interacting constraints between pieces of information. The process of coherence maximisation helps one to form the set of preferences from the available complex network of constraints. Further, coherence-based decision making unlike other multi-attribute decision making processes, works with a dynamic system where everything from beliefs to goals and actions are subject to be selected or discarded. In other theories decision making is more about action selection for a pre-established set of goals based on a pre-established set of criteria.

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

3 Coherence framework

In this section we introduce the generic coherence framework together with those computable functions that will allow us to build coherence-based agents (see Section 5). Our framework is based on Thagrad'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 [7, 24] 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 mind 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. Thagard in his characterisation classifies coherence into different types such as *explanatory, deductive, perceptual, conceptual, analogous* and *deliberative coherence* depending on this underlying relation. Thus, we have different ζ functions for different types of coherence. 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. A guideline for defining ζ is Thagard's guiding principles for each type of coherence, which we will elaborate in Section 4 for deductive coherence.

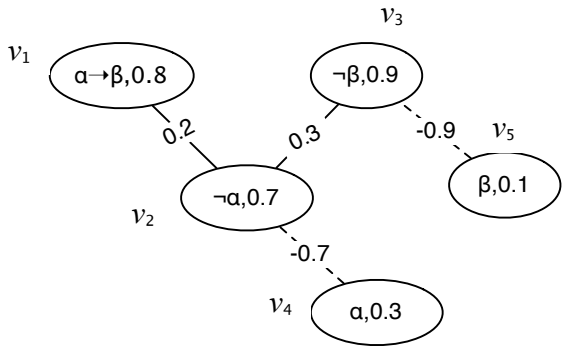

FIG. 1: Graph representing the coherence and incoherence relations between graded propositions related through Modus Tollens: $(\alpha \rightarrow \beta), \neg\beta \vdash \neg\alpha$

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 is constructed with one of the inference rules of the propositional calculus, namely Modus Tollens: $(\alpha \rightarrow \beta), \neg\beta \vdash \neg\alpha$. As we gradually build our framework, we also add more sophistication to our coherence graph in this example.

Thus a coherence graph is defined as follows:

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. $E \subseteq \{\{v, w\} | v, w \in V\}$ is a finite set of edges representing the coherence or incoherence between pieces of information.

3. $\zeta : E \to [-1, 1] \setminus \{0\}$ is an edge-weighted function that assigns a value to the coherence between pieces of information, and which we shall call a *coherence function*

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

Figure 1 is an example of a coherence graph as defined above with the following values.

- $V = \{v_1, v_2, v_3, v_4, v_5\}$
- $E = \{\{v_1, v_2\}, \{v_3, v_2\}, \{v_2, v_4\}, \{v_3, v_5\}\}$
- $\zeta(\{v_1, v_2\}) = 0.3, \zeta(\{v_2, v_4\}) = -0.7, \ldots$

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 $(\mathcal{A}_1, V \setminus \mathcal{A}_1)$ as in Figure 2. We see that, given this partition, the only satisfied constraints are those between $\{v_1, v_2\}$ and between $\{v_2, v_4\}$.

Now we define both the accepted set of the partition that maximises the satisfaction of constraints and the actual value of coherence for this partition. We define first the *strength of a partition* as the sum over the strengths of all the satisfied constraints (ζ values) of that partition. Then the coherence of a graph is defined to be the maximum among the total strengths when calculated over all its partitions. We have the following definitions:

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

$$\sigma(g, \mathcal{A}) = \frac{\sum_{\{v,w\} \in C_{\mathcal{A}}} \mid \zeta(\{v, w\}) \mid}{\mid E \mid} \tag{3.1}$$

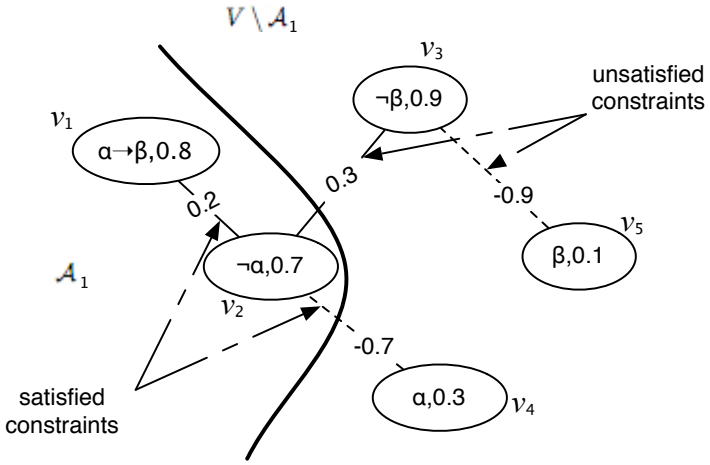

FIG. 2. The strength of partition $(\mathcal{A}_1, V \setminus \mathcal{A}_1)$ is 0.225

For the partition in Figure 2, its *strength* is 0.225.

DEFINITION 3.4
Given a coherence graph $g = \langle V, E, \zeta \rangle$ and given the strength $\sigma(g, \mathcal{A})$ for all subsets $\mathcal{A}$ of V, *the coherence of* g is given by

$$\kappa(g) = \max_{\mathcal{A} \subseteq V} \sigma(g, \mathcal{A}) \tag{3.2}$$

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

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

Thagard's Principle 3 (more discussion on these principles are in Section 4) states that *Propositions that are intuitively obvious have a degree of acceptability on their own.* These obvious propositions are in many cases well proven known facts about context of interest. Hence, we can choose our accepted set to be the one which includes these obvious propositions. However, one can argue that this does not guarantee uniqueness. Another factor to differentiate the accepted sets is the coherence of the sub-graphs restricted to the accepted sets i.e., $g|_{\mathcal{A}}$. The coherence of the sub-graphs gives us an indication of how strongly connected they are. The higher the coherence, the better connected the pieces of information within the sub-graph. Hence, we should prefer an accepted set corresponding to the sub-graph with a higher coherence to that of a subgraph with a lower coherence. Yet another factor to chose an accepted set is the number of nodes in each set. We should prefer those sets with more nodes as we

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

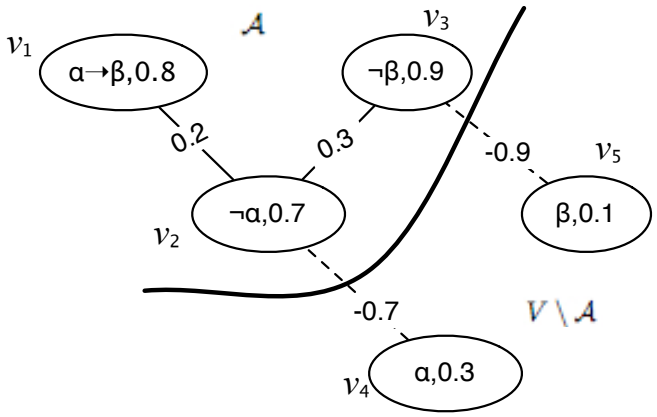

FIG. 3. Coherence of the graph is 0.525 (for the partition $(\mathcal{A}, V \setminus \mathcal{A})$)

For the example in Figure 1, we have a coherence maximising partition $(\mathcal{A}, V \setminus \mathcal{A})$ as in Figure 3. With this partition we see that all the constraints are satisfied and this partition gives the maximum strength for the 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 their associated confidence degrees, we need to define a coherence function ζ relating them. As the nature of relationship between two pieces of information can vary greatly, we do not have one unique coherence function. Thagard in his characterisation of coherence defines different types of coherence namely *explanatory, deductive, conceptual, deliberative, analogic, and perceptual* [29], based on the type of pieces of information and their relationships. Further, in each of of these types, only the corresponding relationship is evaluated. That is, in an explanatory coherence, two pieces of information are coherent only if they are related by an explanation. Thagard proposes certain principles to characterise coherence in each of the different types.

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

adherence with Thagard's principles and later analyse this function in the context of structural and internal connectives. The latter helps us to further derive coherence values between those pieces of information that are not directly related by deduction.

Thagard introduces in [29] the notion of deductive coherence by means of a set of principles:

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

In this section we give a proof-theoretical formalisation of the notion of deductive coherence inspired by the principles put forth by Thagard. We base our coherence function on multiset deductive relations. The concept of a multiset is a generalisation of the concept of a set. Intuitively speaking, we can regard a multiset as a set in which the number of times each element occurs is significant, but not the order of the elements. The introduction of multisets in our framework will allow us to deal more adequately with logics such as linear logics, relevance logics or multi-valued logics. We denote a "multiset deductive relation" as MDR. We assume that all MDRs we deal with are finitary and decidable. These MDRs are often called *simple consequence relations* [3]. We define an MDR as follows:

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

1. **Reflexivity**: $\Gamma \vdash \Gamma$, for every $\Gamma \subseteq L$
2. **Transitivity**: if $\Gamma_1 \vdash \Sigma_1, \gamma$ and $\gamma, \Gamma_2 \vdash \Sigma_2$, then $\Gamma_1, \Gamma_2 \vdash \Sigma_1, \Sigma_2$.

As usual in sequent calculi, we denote by $\vdash \beta$ the fact that 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

Before going into the details of Thagard's principles, it is important to note that these principles are proposed keeping in mind a context or in logical terminology, a theory. Examples of such theories may be 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 rigourous we call it a finite theory presentation $\mathcal{T}$ (however, to avoid lengthy phrases, we also use the term theory to refer to $\mathcal{T}$). 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 information, β (a new norm, a new belief, $\cdots$), it tries to relate this 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 deductive relation with some propositions in $\mathcal{T}$, however, contradicts some other.*

The coherence function we propose here is in the context of a theory and is motivated to aid this process of theory revision. We use Thagard's principles to relate an MDR and the coherence function ζ. Principles 2 and 3 captures the fact that, in a relevant deduction, there are certain positive coherence relations between premises and between each of the premises and the conclusion. Note that, we call only those deductions aided by the theory as relevant. Principle 4 gives an indication of the magnitude of coherence. It states that the magnitude of coherence decreases with the increasing number of premises. Principle 5 discusses the case of contradiction. And finally Principle 7 stresses the basic notion of coherence, that if anything is accepted, it is because, accepting it improves the coherence of the system. Hence, the theory presentation $\mathcal{T}$ is also part of our coherence graph, and its acceptance is only with respect to coherence maximisation.

However, Thagard's principles have been thought about in a boolean world, whereas, in this paper, we define coherence graphs in a more general setting where the propositions can come from either a boolean logic or a many valued logic. Hence, we need to interpret Thagard's principles appropriately. That is, our deductions are graded with the value of the deduction being the grade on the conclusion which is in the range from $[0, 1]$. Since, coherence values lie in the range $[-1, 1]$, we need to normalise the values of the graded deduction, we do so by adding a degree of contradiction, which would in a broad sense measure the degree of the inconsistency between propositions, and ranges between $[-1, 0]$. Coherence function is a summation of these two measures, pushing the values to the positive interval when deduction relation is dominant over the inconsistency relation , and when a contradiction is more significant, deduction giving a value close to zero. Further, to capture Principle 4 we also divide the summation by the number of additional propositions needed to relate the two propositions.

We formalise these with the help of two functions η and ζ, the first capturing the above intuitions, and the second, capturing the fact that coherence is a symmetric function(Principle 1), and hence, the value of coherence between two propositions α and β is the maximum of values of $\eta(\alpha, \beta)$ and $\eta(\beta, \alpha)$. We have the formal definitions below.

Now we define a truth function with respect to a logic. This function helps us to compute the coherence values between formulas by computing the implication and conjunction degrees between formulas.

DEFINITION 4.2
Let $\mathcal{T}$ theory in a language L and is closed under atomic formulas. Then a truth function ρ on L is given by
For $\alpha, \beta \in L$,

- $\rho(\alpha \wedge_L \beta) = F_{\wedge_L}(\rho(\alpha), \rho(\beta))$
- $\rho(\neg_L \alpha) = F_{\neg_L}(\rho(\alpha))$
- $\rho(\bar{0}) = 0$ and $\rho(\bar{1}) = 1$

where $F_{\wedge_L}$ and $F_{\neg_L}$ are the truth connectives defined for L.

We formalise Thagard's principles in terms of a *support function* η on the MDR for a finite theory presentation $\mathcal{T}$ as below.

DEFINITION 4.3
Let L be the set of all propositional sentences of a multi-valued propositional logic. Let $\mathcal{T} \subseteq L$ be a finite theory presentation and $\Gamma \subseteq \mathcal{T}$ and $\gamma \in L$. A support function $\eta_{\mathcal{T}} : L \times L \to [-1, 1]$ with respect to $\mathcal{T}$ is given by

$$\eta_{\mathcal{T}}(\alpha, \beta) = \begin{cases} \max \left\{ \begin{array}{l} \max \left\{ \frac{\rho(\beta)+(F_{\wedge_L}(\rho(\alpha),\rho(\beta))-1)}{|\Gamma|} | \exists \Gamma \subseteq \mathcal{T} : \Gamma, \alpha \vdash \beta \text{ and } \alpha \nvdash \beta \right\}, \\ \max \left\{ \frac{\rho(\gamma)+(F_{\wedge_L}(\rho(\alpha),\rho(\beta))-1)}{|\Gamma|+1} | \exists \Gamma \subseteq \mathcal{T} : \Gamma, \alpha, \beta \vdash \gamma \text{ and } \alpha, \beta \nvdash \gamma \right\} \end{array} \right\} \\ \\ \text{undefined} \qquad \text{otherwise} \end{cases}$$

We now define the deductive coherence between two propositions as the value of the stronger relation since deductive coherence is a symmetric function. Due to this, even if there may only be a deductive relation in one direction, there will be a deductive coherence in both directions. 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.

Then we have the coherence function defined as follows:

DEFINITION 4.4
Let L be the set of all propositional sentences of a multi-valued propositional logic. Let $\mathcal{T} \subseteq L$ be a finite theory presentation and let $\eta_{\mathcal{T}} : L \times L \to [-1, 1]$ be a support function. A *deductive coherence function* $\zeta_{\mathcal{T}} : L \times L \to [-1, 1] \setminus \{0\}$ with respect to $\mathcal{T}$ is a partial function given by:

For any pair (α, β) of formulas in L,

$$\zeta_{\mathcal{T}}(\alpha, \beta) = \begin{cases} \max(\eta_{\mathcal{T}}(\alpha, \beta), \eta_{\mathcal{T}}(\beta, \alpha)) & \text{if } \eta_{\mathcal{T}}(\alpha, \beta) \text{ and } \eta_{\mathcal{T}}(\beta, \alpha) \text{ are defined and } \neq 0 \\ \\ \eta_{\mathcal{T}}(\alpha, \beta) & \text{if } \eta_{\mathcal{T}}(\alpha, \beta) \text{ defined and } \neq 0 \\ & \text{and } \eta_{\mathcal{T}}(\beta, \alpha) \text{ undefined or } = 0 \\ \\ \eta_{\mathcal{T}}(\beta, \alpha) & \text{if } \eta_{\mathcal{T}}(\beta, \alpha) \text{ defined and } \neq 0 \\ & \text{and } \eta_{\mathcal{T}}(\alpha, \beta) \text{ undefined or } = 0 \\ \\ \textit{undefined} & \text{if } \eta_{\mathcal{T}}(\alpha, \beta) \text{ and } \eta_{\mathcal{T}}(\beta, \alpha) \text{ are undefined or } = 0 \end{cases}$$

In our example (Figure 1), let us further assume that

$$\mathcal{T} = \{\alpha \to \beta, \neg\beta \to \alpha\}$$

Then we have

$$\eta(\neg\beta, \neg\alpha) = \frac{\rho(\neg\alpha) + F_{\wedge_L}(\rho(\neg\beta), \rho(\neg\alpha)) - 1}{|\Gamma|} = \frac{0.7 + (0.6 - 1)}{1} = 0.3$$

Since this is the only deduction relation between these formulas in the context of $\mathcal{T}$ $\zeta(\{\neg\beta, \neg\alpha\}) = 0.3$. Similarly, since $\eta(\alpha, \neg\alpha) = \frac{0.3+(-1)}{1} = -0.7$, we also have $\zeta(\{\alpha, \neg\alpha\}) = -0.7$.

4.2 Reduction to a Two-valued Logic

Since Thagard's principles are based on a two-valued logic, we now reduce our general definitions of coherence functions to for a two-valued logic, and show that they satisfy Thagard's Principles. Further we also prove in the next section certain properties of the coherence function by exploiting some of the structural rules and connectives of the logic. We prove these properties staying in a boolean world. To prove similar properties for graded logic is one of our future works.

Definition 4.5
Let L be the set of all propositional sentences of a two-valued propositional logic. Let $\mathcal{T} \subseteq L$ be a finite theory presentation and $\Gamma \subseteq \mathcal{T}$. A support function $\eta_{\mathcal{T}} : L \times L \to [-1, 1]$ with respect to $\mathcal{T}$ is given by

$$\eta_{\mathcal{T}}(\alpha, \beta) = \begin{cases} \max \left\{ \begin{array}{l} \max \left\{ \frac{1}{|\Gamma|} | \exists \Gamma \subseteq \mathcal{T} : \Gamma, \alpha \vdash \beta \text{ and } \alpha \nvdash \beta \right\}, \\ \max \left\{ \frac{1}{|\Gamma|+1} | \exists \Gamma \subseteq \mathcal{T} : \Gamma, \alpha, \beta \vdash \gamma \text{ and } \alpha, \beta \nvdash \gamma \right\} \end{array} \right\} \\ \frac{-1}{|\Gamma|} \qquad \text{if } \Gamma, \alpha, \beta \vdash \\ \text{undefined} \qquad \text{otherwise} \end{cases}$$

Then we have the coherence function for a two valued propositional logic as follows:

Definition 4.6
Let L be the set of all propositional sentences of a two-valued propositional logic. Let $\mathcal{T} \subseteq L$ be a finite theory presentation and let $\eta_{\mathcal{T}} : L \times L \to [-1, 1]$ be a support function. A *deductive coherence function* $\zeta_{\mathcal{T}} : L \times L \to [-1, 1] \setminus \{0\}$ with respect to $\mathcal{T}$ is a partial function given by:

For any pair (α, β) of formulas in L,

$$\zeta_{\mathcal{T}}(\alpha,\beta) = \begin{cases} \max(\eta_{\mathcal{T}}(\alpha,\beta), \eta_{\mathcal{T}}(\beta,\alpha)) & \text{if } \eta_{\mathcal{T}}(\alpha,\beta) \text{ and } \eta_{\mathcal{T}}(\beta,\alpha) \text{ defined and } \neq 0 \\ \eta_{\mathcal{T}}(\alpha,\beta) & \text{if } \eta_{\mathcal{T}}(\alpha,\beta) \text{ defined and } \neq 0 \\ & \text{and } \eta_{\mathcal{T}}(\beta,\alpha) \text{ undefined or } = 0 \\ \eta_{\mathcal{T}}(\beta,\alpha) & \text{if } \eta_{\mathcal{T}}(\beta,\alpha) \text{ defined and } \neq 0 \\ & \text{and } \eta_{\mathcal{T}}(\alpha,\beta) \text{ undefined or } = 0 \\ \textit{undefined} & \text{if } \eta_{\mathcal{T}}(\alpha,\beta) \text{ and } \eta_{\mathcal{T}}(\beta,\alpha) \text{ are undefined or } = 0 \end{cases}$$

PROPOSITION 4.7
A deductive coherence function $\zeta_{\mathcal{T}}$ satisfies Thagards's principles introduced in the beginning of this section.

PROOF.
Principle1 : $\zeta_{\mathcal{T}}$ is symmetric by construction.
Principle 2 : Given $\Gamma \subseteq \mathcal{T}$ with $|\Gamma| = n$ and $\alpha \in \mathcal{T}$ and if $\Gamma, \alpha \vdash \beta$ and there is no other deduction relating α and β for any $|\Gamma'| < |\Gamma|$ then $\zeta_{\mathcal{T}}(\{\alpha, \beta\}) = \frac{1}{n}$.
Principle 3: Given $\Gamma \subseteq \mathcal{T}$ with $|\Gamma| = n$ and $\alpha, \beta \in \mathcal{T}$ and if $\Gamma, \alpha, \beta \vdash \gamma$ and there is no other deduction relating α and β for any $|\Gamma'| < |\Gamma|$ then $\zeta_{\mathcal{T}}(\{\alpha, \beta\}) = \frac{1}{n+1}$.
Principle 4 :Given $\Gamma_1, \Gamma_2 \subseteq \mathcal{T}$ with $|\Gamma_1| = n$ and $|\Gamma_2| = m$, $n \leq m$ and $\alpha \in \mathcal{T}$ and if $\Gamma_1, \alpha_1 \vdash \beta$ and $\Gamma_2, \alpha_2 \vdash \beta$ and there is no other deduction relating α_1 and β for any $|\Gamma'| < |\Gamma_1|$ and that relating α_2 and β for any $|\Gamma'| < |\Gamma_2|$ then $\zeta_{\mathcal{T}}(\{\alpha_1, \beta\}) = \frac{1}{n} > \zeta_{\mathcal{T}}(\{\alpha_2, \beta\}) = \frac{1}{m}$.
Principle 5 : Satisfied by construction.
Principle 6 :Propositions that are intuitively obvious are the axioms of the theory. That is, if $\alpha \to \beta$ is an axiom in $\mathcal{T}$, then we have $\zeta_{\mathcal{T}}(\{\alpha, \beta\}) = 1$. That is, β coheres with α with the highest degree. Hence β has an intuitive priority. The more this axiom is used in deriving other propositions, the higher the chance of this axiom being accepted.
Principle 7 : Satisfied by the Equation 3.2. ∎

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

PROPOSITION 4.8
Whenever $\alpha, \beta, \gamma \nvdash$ and $\nvdash \alpha$, $\nvdash \beta$, $\nvdash \gamma$,

$$\zeta_{\mathcal{T}}(\{\gamma, \alpha\}) = 1 \text{ and } \zeta_{\mathcal{T}}(\{\alpha, \beta\}) = 1, \text{ then } \zeta_{\mathcal{T}}(\{\gamma, \beta\}) = 1$$

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

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

4.3 Properties of Coherence Based On MDRs

We can classify logics according to structural rules (such as weakening /monotonicity) and connectives available in it. There are two types of connectives: the *internal* connectives, which transform a given sequent into an equivalent one that has a special required form, and the *combining* connectives, which combine two sequents into one. For instance, classical propositional logic is monotonic, satisfies weakening, has the internal and combining connectives, and makes no difference between the combining and the corresponding internal connectives. On the other hand, propositional linear logic is monotonic, has the above connectives but distinguishes between internal and combining ones. Intuitionistic logic differs from classical propositional logic in its implication connective and does not contain any internal negation. In this section, we explore the properties of the deductive coherence function $\zeta_{\mathcal{T}}$ which would help determine the value of function $\zeta_{\mathcal{T}}$[2] between pairs of formulas which are related through some of the structural rules and connectives. We do this by identifying the properties of the support function $\eta_{\mathcal{T}}$ using the properties of the connectives and structural rules.

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

4.3.1 Combining Conjunction

Conjunctive $\wedge$ is *combining* iff, for all $\Gamma \subseteq \mathcal{T}$, and $\Sigma \subseteq L$ and $\alpha, \beta \in L$ where $\mathcal{T} \subseteq L$ is a theory of the language L we have

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

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

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

4.3.2 Internal Conjunction

Conjunction $\circ$ is *internal* iff, for all $\Gamma \subseteq \mathcal{T}$, and $\Sigma \subseteq L$ and $\alpha, \beta \in \mathcal{T}$ where $\mathcal{T} \subseteq L$ is a theory of the language L we have

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

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

[2] From now on we drop the suffix for convenience for both $\zeta_{\mathcal{T}}$ and for $\eta_{\mathcal{T}}$, however it should be noted that they are always evaluated with respect to a finite theory presentation $\mathcal{T}$.

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

4.3.3 Combining Disjunction

Disjunction $\vee$ is *combining* iff, for all $\Gamma \subseteq \mathcal{T}$, and $\Sigma \subseteq L$ and $\alpha, \beta \in \mathcal{T}$ where $\mathcal{T} \in L$ we have

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

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

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

4.3.4 Internal Disjunction

Disjunction $+$ is *internal* iff, for all $\Gamma \subseteq \mathcal{T}$, and $\Sigma \subseteq L$ and $\alpha, \beta \in L$ where $\mathcal{T} \subseteq L$ is a theory of the language L we have

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

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

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

4.3.5 Combining Implication

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

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

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

1. Given that $\Gamma, \alpha \supset \beta \vdash \sigma$ implies $\Gamma, \beta \vdash \sigma$ we have that
   if $\eta(\alpha \supset \beta, \sigma) = \frac{1}{n}$ then $\eta(\beta, \sigma) = \frac{1}{n}$
2. Given that $\gamma, \alpha \supset \beta \vdash$ iff $\gamma \vdash \alpha$ and $\gamma, \beta \vdash$, we have that
   if $\eta(\gamma, \alpha \supset \beta) = -1$ iff $\eta(\gamma, \alpha) = 1$ and $\eta(\gamma, \beta) = -1$
3. Given $\alpha, \beta \nvdash$ then we have that
   $\eta(\beta, \alpha \supset \beta) = 1$
   and if $\vdash$ satisfies weakening, then we have that
   $\eta(\alpha, \beta) = 1$ and $\eta(\alpha \supset \beta, \beta) = 1$

4.3.6 Internal Implication

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

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

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

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

4.3.7 Internal Negation

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

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

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

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

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

4.4 An Example

So far we have analysed formally the properties of coherence by listing the properties of the support function in terms of the connectives and structural properties of a logic. Here we apply these properties in the context of many valued propositional logic to deduce some of the coherence values. We use the same example as in the previous sections. We enrich the example by adding another implication namely $\gamma \vdash \alpha \rightarrow \beta$ to the theory $\mathcal{T}$. Hence,

$\mathcal{T} = \{\alpha \rightarrow \beta, \neg\beta \rightarrow \alpha, \quad \gamma \vdash \alpha \rightarrow \beta, \quad \gamma, \neg\beta \rightarrow \alpha\}$

Then using the properties we have derived so far, we can deduce the following coherence values. Since in the example we only have implications and negations, we use only the properties related to implication and negation. It is however easy to see how we can similarly apply the results of other connectives in appropriate cases. The purpose of this example is only to demonstrate that coherence graphs can be enriched with these properties of coherence, however, we do not intend to be exhaustive.

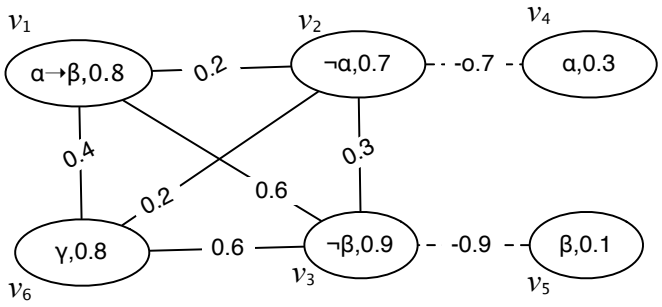

FIG. 4. Applying properties of η to compute coherence values

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 which 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 the BDI theory [26], since it is one of the prominent existing agent architectures. We use an adaptation of the architecture developed by Casali et al. [9] based on multi-context systems (MCS), which incorporate graded cognitions. The grade in a cognition represents the degree to which an agent believes (desires or intents) a particular cognition. We use graded cognitions to incorporate reasoning under uncertainty into our agent framework. Then, an MCS models the representation and interaction between these graded cognitions.

In the work of Casali et al., the MCS specification of an agent contains three basic components: units or contexts, logics, and bridge rules, which channel the propagation of consequences among the 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 rules $\vdash_i$ inside each context, and bridge rules B between contexts. Internal rules allow an agent to draw consequences within a context, while bridge rules allow to embed results from one context into another [17, 18]. Thus, an agent is defined as a family of interconnected contexts:

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

where

- each context $C_i = \langle L_i, A_i, \vdash_i, T_i \rangle$ consists of a language L_i, a set of axioms A_i, and a deductive relationship $\vdash_i$. They define the logic for the context and its basic behaviour as constrained by the axioms. In addition a theory $T_i \in L_i$ is associated with each context, which represent the particular axioms of the context.
- B, the set consists of inference rules with premises and conclusions in different contexts.
  For instance:
  $$\frac{1:\psi, 2:\varphi}{3:\phi}$$
  represents that if formula ψ is deduced in context C_1 and formula φ is deduced in context C_2 then formula ϕ is inferred in context C_3.

The multi-context architecture is adapted here as we add further structure by associating a coherence graph to the theory of each of the contexts. In an MCS, internal rules $\vdash_i$ allow to draw consequences within a theory, while bridge rules allow to embed results from a theory into another. Since we have coherence graphs associated with the theories in the contexts, our bridge rules carry consequences from one graph to another. In the following we define the belief, desire, intention and norm contexts corresponding to the normative BDI agent we are interested in. We also define how an extension of the bridge rules can be used to reason across these contexts. Once we have these components defined, we discuss the agent architecture itself.

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. [8]. For the norm logic associated with the norm context, we use the the work of Godo et al [13] on probabilistic deontic logic. We here give a sketch of a belief context, while the details are in Casali et.al [9]. The desire, intention, and norm contexts can be defined similar to the definition of the belief context, with the belief logic replaced by the desire, intention or norm logic. As in Casali et.al, we define first a logic for the context, in this case a belief logic, consisting of a belief language, a set of axioms and a deductive relationship. Further, we extend the belief context by associating a belief coherence graph, whose nodes are formulas of the belief language. In particular, these graphs are deductive coherence graphs, as we have only defined a deductive coherence function so far.

A belief logic $\mathcal{K}_B$ consists of a belief language, a set of axioms and a deductive relation defined on the belief logic. $\langle L_B, A_B, \vdash_B \rangle$. We define the belief language L_B by extending the classical propositional language L defined upon a countable set

of propositional variables PV and connectives $(\neg, \rightarrow)$. We extend L with a fuzzy unary modal operator B. The modal language L_B is built from the elementary modal formulae $B\varphi$ where φ is propositional, and truth constants r, for each rational $r \in Q \cap [0,1]$, 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 $\mathcal{K}_B$[3]. Formally the belief language L_B is defined as:

Definition 5.1
[8] Given a propositional language L, a belief language L_B is given by:

- If $\varphi \in L$ then $B\varphi \in L_B$
- If $r \in Q \cap [0,1]$ then $\overline{r} \in L_B$
- 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 the conjunction and implication of Łukasiewicz logic)

Other Łukasiewicz logic connectives for the modal formulae can be defined from &, $\rightarrow_L$ and $\overline{0}$: $\neg_L\Phi$ is defined as $\Phi \rightarrow_L \overline{0}$, $\Phi \wedge \Psi$ as $\Phi \& (\Phi \rightarrow_L \Psi)$, $\Phi \vee \Psi$ as $\neg_L(\neg_L\Phi \wedge \neg_L\Psi)$, and $\Phi \equiv \Psi$ as $(\Phi \rightarrow_L \Psi) \& (\Psi \rightarrow_L \Phi)$. Since in Łukasiewicz logic a formula $\Phi \rightarrow_L \Psi$ is 1-true iff the truth value of Ψ is greater or equal to that of Φ, modal formulae of the type $\overline{r} \rightarrow_L B\varphi$ express that the probability of φ is at least r. Formulae of the type $\overline{r} \rightarrow_L \Psi$ will be denoted as (Ψ, r).
We call $\mathcal{T}_B$ a theory of the language L_B.
The axioms A_B of $\mathcal{K}_B$ are:

1. All axioms of propositional logic.
2. Axioms of Łukasiewicz logic for modal formulas (for instance, axioms of Hájek's Basic Logic (BL) [19] plus the axiom: $\neg\neg\Phi \rightarrow \Phi$.)
3. Probabilistic axioms, given $\varphi, \psi \in L$:
   - $B(\varphi \rightarrow \psi) \rightarrow_L (B\varphi \rightarrow B\psi)$
   - $B\varphi \equiv \neg_L B(\varphi \wedge \neg\psi) \rightarrow_L B(\varphi \wedge \psi)$

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

1. Modus ponens.
2. Necessitation for B (from φ derive $B\varphi$).

Note that the truth function $\rho : L_B \rightarrow [0,1]$ is defined by means of the truth-functions of Łukasiewicz logic and the probabilistic interpretation of beliefs as follows:

- $\rho(B\varphi, r) = r$ for all $r \in Q \cap [0,1]$
- $\rho(\varphi \,\&\, \psi) = max(\rho(\varphi) + \rho(\psi) - 1, 0)$ for all $\varphi, \psi \in L_B$
- $\rho(\varphi \rightarrow_L \psi) = min(1 - \rho(\varphi) + \rho(\psi), 1)$ for all $\varphi, \psi \in L_B$

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

Definition 5.2
Given a belief logic $\mathcal{K}_B = \langle L_B, A_B, \vdash_B \rangle$ where L_B is a belief language, A_B are a set of axioms and $\vdash_B$ are a set of deduction rules, a belief graph $g_B = \langle V_B, E_B, \zeta_B \rangle$ is a coherence graph defined over $\vdash_B$ and a finite theory $\mathcal{T}_B$ of L_B such that:

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

- $V_B \subseteq \mathcal{T}_B$
- E is a set of subsets of 2 elements of V_B
- ζ_B is the deductive coherence function defined over $\vdash_B$ and $\mathcal{T}_B$.

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

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

5.1.1 Norm Graph (g_N)

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

In order to define a *norm graph*, we need to first define a norm logic $\mathcal{K}_N = \langle L_N, A_N, \vdash_N \rangle$. As we have graded norms, we define $\mathcal{K}_N$ as a graded deontic logic namely the Probability-valued Deontic Logic [13] to represent and reason with norms. We define the norm language L_N by extending the classical propositional language L defined upon a countable set of propositional variables and connectives $(\neg, \rightarrow)$. L_N is defined as a fuzzy modal language over Standard Deontic Logic (SDL) to reason about the probability degree of deontic propositions. In our case the probability values are replaced by the grades associated with the norms. The language, axioms and deductions rules are defined similarly as in the case of the belief logic. For the details, refer [13].

Here we list some of the examples of valid norms in PSDL. Below we use triples form for propositional formulas. Also, to keep uniformity with the cognitive languages, we adopt a slightly different notation from that given in [13].

- $(O\langle John, use, public\_transport\rangle \rightarrow \langle John, validate, ticket\rangle, 0.8)$
  If John uses public transport, then John is obliged to validate the ticket.
- $(O\langle Anna, citizen, Utopia\rangle \rightarrow \langle Anna, pay\_tax, Utopia\rangle, 1)$
  If Anna is a citizen of Utopia, then Anna ought to pay tax to Utopia.

5.2 Reasoning Across Contexts

Reasoning in a BDI normative agent needs to consider the influence of cognitions and norms among each other. For instance, it is desirable to chose or predict an action that is most coherent with the set of cognitions and accepted norms. It is also desirable to know whether a new information is coherent with the set of existing information, and if not, whether a cognitive revision is desirable given the new information or a revision

of commitment to a norm. Typically, in a multi-context system, such reasoning is achieved by reasoning across contexts through a set of bridge rules. Since, in a coherence-driven agent, theories of each context are nodes of a coherence graph, the agents need to be able to reason across coherence graphs.

Bridge rules are in certain sense inference rules, which carry inferences between theories of different logics. Since, the theories are nodes of individual coherence graphs, we can use these inferences 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. That is, we define two functions, an extension function that conceptually extend certain theories by adding new formulas (conclusions of the inferences). In a graph terminology, this amounts to adding new nodes and thus extending some of the graphs participating in the inference. Further, there exists positive coherence relation between formulas of the premises an conclusion of the inference rules. Thus we join the coherence graphs by adding edges between formula nodes of premises and conclusion. Below we define both the graph extension and edge extension functions and finally, we discuss the definition and application of these functions when the inference rules are bridge rules.

A graph node extension function (denoted with ε) takes into account the influence of graphs (theories) on each other. Let's assume, for instance, that an agent wants it to be the case that whenever it has an intention $(I\varphi, r)$ in the intention graph (a formula in the theory $\mathcal{T}_I$), then the corresponding belief $(B\varphi, r)$ is inferred in the belief graph (added to the theory $\mathcal{T}_B$).

Definition 5.3
Given $n > 0$, we say that a function $\varepsilon : \mathcal{G}^n \to \mathcal{G}^n$ is a *graph extension function* if, given a tuple of graphs $\bar{g} = \langle g_1, \cdots, g_n \rangle$ in $\mathcal{G}^n$, $\varepsilon(\bar{g}) = \bar{g}'$ is such that

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

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

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

Definition 5.4
Given $n, j > 0$, we say that a sequence is an *extension sequence* if, given a tuple of graphs $\bar{g} \in \mathcal{G}^n$ and a set of functions $S \subseteq \mathcal{E}$,

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

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

An edge extension function (denoted with ι) joins a set of graphs by adding edges between the nodes participating in the inference. Consider our previous example and let's assume that our agent further wants it the case that, the belief and the intention

nodes are related and have a positive coherence between them. Note that this does not change the theories, as we are only making new associations between members of different theories. The function ι takes n graphs and joins the graphs by adding new edges (and coherence values on the edges) between related nodes. We have the following definition for ι:

Definition 5.5
Given $n > 0$, we say that a function $\iota : \mathcal{G}^n \to \mathcal{G}$ is a *graph join function* if given a tuple of graphs $\bar{g}$ in $\mathcal{G}^n$, $\iota(\bar{g}) = \langle V, E, \zeta \rangle$ such that:

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

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

Now we define the composition of graphs in a tuple $\bar{g}$ by combining the two functions ε and ι. That is, we apply the set of functions $T \subseteq \mathcal{J}$ on the fixed point of the set of functions $S \subseteq \mathcal{E}$ applied over $\bar{g}$. The union of all the resulting graphs is defined as a composition of graphs. Note that here we assume S has a unique fixed point applied over any tuple of graphs $\bar{g}$. It is however a fair assumption given that we can construct the functions in S and T according to the requirements. Further, it should be noted that we keep the theories separate and only compose the corresponding coherence graphs.

Definition 5.6
Given $n > 0$, we say that a function $\varsigma : \mathcal{G}^n \to \mathcal{G}$ is a *graph composition function* if, given a tuple of graphs $\bar{g}$ in $\mathcal{G}^n$, a set of functions $S \subseteq \mathcal{E}$ with a unique fixed point and a set of functions $T \subseteq \mathcal{J}$ then $\varsigma(\bar{g}) = \bigcup_{\iota \in T} \iota(S^*(\bar{g}))$.

5.2.1 Bridge Rules — A Set of Composition Functions

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

Definition 5.7
Given $n > 0$, a bridge rule b is a rule of the form

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

with:

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

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

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

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

Definition 5.8
Given a tuple of graphs $\bar{g} = \langle g_1, g_2, \cdots, g_n \rangle$, a bridge rule

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

as in Definition 5.7 and if, for all k we have $(\pi(A_k), r_k) \in V_k$, where π is the most general substitution making the formula schemata A_k match nodes in V_k, then an extension function ε and a join function ι are derived from b as:

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

2. $\iota(\bar{g}) = \langle V, E, \zeta \rangle$ as in Definition 5.5 such that:
   For all $1 \leq k \leq q$, and $1 \leq m \leq q$:
   - $V = \bigcup_{i \neq j} V_i$
   - $\{\pi(A_k), \pi(A)\} \in E$;
   - $\{\pi(A_k), \pi(A_m)\} \in E$;
   - $\eta(\pi(A_k), \pi(A)) = q + 2$;
   - $\eta(\pi(A_k), \pi(A_m)) = q + 2$
   - $\zeta(\{v, w\}) = 1/min\{\eta(v, w), \eta(w, v)\}$ for all $v, w \in V$

5.2.2 Application of Composition Functions — An Example

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

1. Given a bridge rule
$$b = \frac{1 : (B\psi, r), 2 : (D\psi, s)}{3 : (I\psi, \min(r, s))}$$
   Where the indices $1, 2$ and 3 correspond to the contexts C_1, C_2, and C_3 which are associated with the graphs g_B, g_D and g_I respectively.
2. And Given $(B\psi, 0.95) \in g_B$, $(D\psi, 0.95) \in g_D$

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

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

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

- $E = E_B \cup E_D \cup E_I \cup \{\{(I\psi, 0.95), (B\psi, 0.95)\}, \{(I\psi, 0.95), (D\psi, 0.95)\}\}$
- $\zeta(\{(I\psi, 0.95), (B\psi, 0.95)\}) = 0.425$;
  $\zeta(\{(I\psi, 0.95), (D\psi, 0.95)\}) = 0.425$

5.3 Coherence-driven Agents

Equipped with the 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, inference rules, and an associated theory respectively. In our extension of MCS, a coherence-driven agent will further have a function f that maps the set of theories of each context to a set of corresponding coherence graphs. And a function h that maps a set of bridge rules to a set of graph composition functions. These extensions are because the theories are expressed as coherence graphs. An agent will need both the coherence functions and a set of deduction mechanisms to reason within and between graphs. For the BDI agents considered here, the contexts are C_1, C_2, C_3 and C_4 which determine a belief graph g_B, a desire graph g_D, an intention graph g_I and a norm graph g_N respectively. Hence we have the following definition:

Definition 5.9
A *coherence-driven agent* a is a tuple $\langle \{C_i\}_{1 \leq i \leq 4}, B, f, h \rangle$ where $\{C_i\}$ is a family of contexts, $B \subseteq \mathcal{B}$ is a set of bridge rules, $f : \{\mathcal{T}_i\} \rightarrow \mathcal{G}$ maps theories of each context to coherence graphs, and $h : B \rightarrow \mathcal{E} \times \mathcal{J}$ maps bridge rules to pairs of graph extension and graph join functions.

In the following we describe how coherence-driven agents interact with a normative environment. An agent at any time can either perceive the normative environment or make a prediction about a future action. The perception from a normative environment can be of two types, the first type consists of a new information in the form of a new belief and the second type in the form of a meta-information, such as a norm (a norm change or a new norm). As we express norms as conditionals (see 5.1.1), a normative society interprets the conclusions of a norm to be obligatory or permissive whenever its premisses hold. For a normative agent, this translates to, whenever it believes that the premisses are true, then it should or can intend the conclusion. Hence, when the observation is a norm, the agent not only should incorporate the norm into its theory, but also the belief on its conditionals and its intention of the conclusion. This is a form of hypothetical reasoning where the agent tries to evaluates the implications of the norm by assuming its premises or in other words, tries to understand what it means to follow the proposed norm. In both the cases, a coherence-driven agent re-evaluates the coherence of the collection of information including the new information. This evaluation enable the agent to know the status

of the new information in terms of its acceptability, while also knowing the relative status of the old information with respect to the new one. In the case, a prediction needs to be made, a coherence-driven agent uses only the set of accepted information it has.

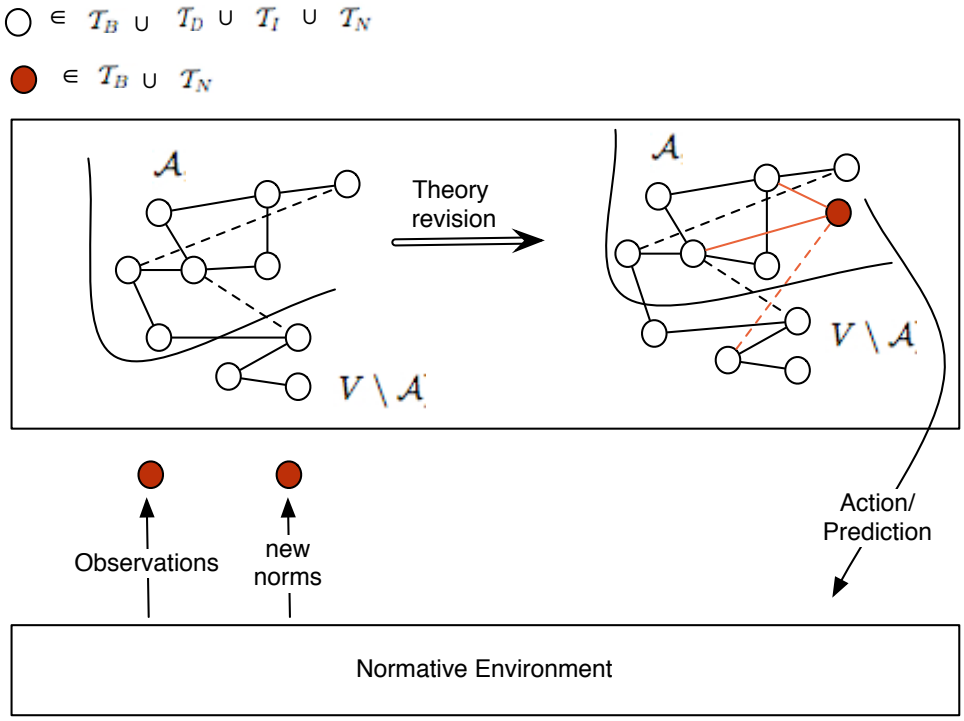

FIG. 5.

As in Figure 5, a coherence-driven agent starts with a set of theories corresponding to its beliefs, desires, intentions and norms it has accepted. It is desirable for this set of theories to be consistent, however, our proposal is tolerant to inconsistencies and in certain sense expose and eliminate them if possible. Further, the agent is assumed to have certain rules to reason across theories. An agent first express its theories as nodes in coherence graphs and computes composite graphs using its rules to reason across theories. It further computes the coherence maximising partition to know its most coherent set of cognitions and norms.

In the event of a new information, an agent re-evaluates its theory, hence recomputes both the composite graphs and the coherence maximising partition. As discussed in Section 4, if the new information reinforces the theory, it is added to the accepted set and the theory becomes more coherent. However, if it contradicts elements of the accepted theory, then either the new information is rejected or some part of the already accepted theory is rejected, whichever makes the theory more coherent. However, to make predictions, it uses only the accepted theory. This is realistic, as it is the accepted set that the agent wants to base its decisions on. In the following

we outline this process in an algorithm, which we will later follow to demonstrate the reasoning of a coherence-driven agent.

Below we describe the procedure a coherence-driven agent follows when it encounters a new observation. We assume the following input for the procedure. To simplify the algorithm, we assume just one bridge rule as part of the input.

- the new observation $(K\varphi, r)$ where $K \in \{B, D, I, O, P\}$, though usually direct observations are either beliefs or norms whereas the other cognitions are often derived from these;
- the cognitive and norm coherence graphs of the agent g_B, g_D, g_I, and g_N;
- a bridge rule $b \in \mathcal{B}$.

ALGORITHM 5.10

1: $v := (K\varphi, r)$
2: **if** $K \in \{O, P\}$ **then**
3: $K := N$
4: **end if**
5: $V_K := V_K \cup \{v\}$
6: **for** all $w \in V_K$ **do**
7: compute $\zeta(\{v, w\})$ using Definition 4.4.
8: **if** $\zeta(\{v, w\})$ is defined **then**
9: $E_K := E_K \cup \{\{v, w\}\}$
10: **end if**
11: **end for**
12: $\bar{g} := \langle g_B, g_D, g_I, g_N \rangle$
13: **while** $\bar{g} \neq \varepsilon(\bar{g})$ **do**
14: $\bar{g} := \varepsilon(\bar{g})$ where ε as in Definition 5.3 derived from bridge rule b
15: $\varsigma_g := \varsigma_g \cup \iota(\bar{g})$ where ι as in Definition 5.5 derived from b and ς_g is the composite graph as in Definition 5.6.
16: **end while**
17: **for** all $(\mathcal{A}_i, V \setminus \mathcal{A}_i)$, $\mathcal{A}_i \subseteq V$ **do**
18: calculate $\sigma(\varsigma_g, \mathcal{A}_i)$ using Equation 3.1
19: **end for**
20: $\kappa := \kappa(\varsigma_g)$ using Equation 3.2
21: $\mathcal{A} := \mathcal{A}_i | \max(\sigma(\varsigma_g, \mathcal{A}_i))$

The lines from 1 to 11 updates the graphs by incorporating the new observation and its influences on existing elements of the observed cognition. When the observation is a new norm, however, there may be additional elements such as an associated belief on sanctions and rewards which is together observed with the norm. In such cases this algorithm is iterated over all the new observations.

Lines from 120 to 16 builds up the reasoning across contexts by composing the coherence graphs. First, the particular graphs are extended by taking into account the influence of the new belief $(B\varphi, r)$. The graph extension function ε is used to extend the graphs. As and when we extend a graph, we also perform the corresponding edge extension using the join function ι. Finally, we incrementally make a composite graph by making a union of the composite graph so far with the most recent join of the graphs. Note that we repeat the above steps until we reach a fixed point, or until e cannot find new graph extensions.

In the second part of the algorithm, lines from 17 to 21 determines the coherence maximising partition. This is done by first computing the strength of each partition using the function σ and choosing the partition $(\mathcal{A}, V \setminus \mathcal{A})$ for which $\sigma(g, \mathcal{A})$ is maximal. This part of the algorithm only gives the simplest solution, however, finding a maximising partition of a weighted graph is known to be an NP-complete problem. There are approximation algorithms exist to find the solution to this problem such as max-cut, neural network based algorithms.

Further, in finding the accepted set using the above algorithm, we make an assumption that there is only one accepted set corresponding to the maximum value of $\sigma(g, \mathcal{A}_i)$. However this is not true, because even if there is just one partition $(\mathcal{A}, V \setminus \mathcal{A})$ corresponding to the maximum, $\mathcal{A}$ and its dual $V \setminus \mathcal{A}$ are already two accepted sets which produces the same maximum value of coherence. In addition there could be other partitions which maximises the coherence. According to the guidelines discussed in 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.

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

6 Example — Norm Evaluation

We apply the formalism developed in the previous sections to model norm evaluation in a real scenario. The example is motivated by the water sharing treaty signed between the southern states of India during 1892 and 1924 and the disputes thereafter [32]. The objectives of this example are threefold. First, to demonstrate how self-interested agents working together evaluate norms. Second, to show the need for *norm adaptation* inspired by individual coherence evaluations, whereas the grander aim is to set up a framework for norm adaptation itself, which will be our future work. And third, to open new application areas in norm evaluation where such cognitive theories could be applied.

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

6.1 Terminology

To represent the cognitions and norms concerning an agent, we shall have belief, desire, intention and norm languages as defined in Section 5.1. Hence, $(B\varphi, r)$ rep-

resents that the agent believes that proposition φ is true (in a near future world[4]) with degree r. (Propositions $(D\varphi, r)$, and $(I\varphi, r)$ are desires and intentions and are interpreted analogously). $(O(\varphi \rightarrow \psi), r)$ $((P(\varphi \rightarrow \psi), r))$ is the obligation (permission) on the agent to make ψ whenever φ is believed to be true. The degree r is a measure on the norm, such as the priority of the norm, or to what extend it needs to be fulfilled. The statements about the world are in propositional language where each proposition is a triple of the form $\langle object, attribute, value\rangle$. For instance $\langle urbanization, growth\_index, high\rangle$ states that *there is a high growth in urbanisation.*

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

1. $b_1 = \frac{3:(I\psi,r)}{2:(B\psi,r)}$: Whenever there is an intention $(I\psi, r)$ in the theory of the context C_3, then a corresponding belief $(B\psi, r)$ inferred in the theory of context C_1.
2. $b_2 = \frac{1:(B\psi,r),2:(D\psi,s)}{3:(I\psi,\min(1-r+s,1))}$: Whenever there is a belief ψ with a degree r in the theory of C_1 and a desire ψ with a degree s in the theory of C_2, then a corresponding intention ψ with a degree $\min(1 - r + s, 1)$ is inferred in the theory of context C_3.
3. $b_3 = \frac{1:(B(\varphi\rightarrow\psi)\rightarrow\delta,r),3:(I(\delta),s)}{4:(O(\varphi\rightarrow\psi),\min(1-r+s,1))}$: If a belief in C_1 that a norm aids in achieving an intention in C_3, then it is obliged in C_4.
4. $b_4 = \frac{1:(B(\varphi\rightarrow\neg\psi)\rightarrow\delta,r),3:(I(\neg\delta),s)}{4:(O(\varphi\rightarrow\psi),\min(1-r+s,1))}$. If a belief in C_1 that not following a norm triggers the contrary of an intention in C_3, then the norm is obliged in C_4.
5. $b_5 = \frac{4:(O(\varphi\rightarrow\psi),r)}{3:(I(\varphi\rightarrow\psi),r)}$. Whenever there is an obligation in C_4 then the corresponding intention is inferred in C_3.

6.2 Norm Adoption

Snapshot t_1 : 1892
Agent : s
New information: A new norm—agent s should release 300 billion ft^3 of water to agent t annually. And an associated sanction of a military action in the case of norm proposal rejected.

$\mathcal{T}_B$	$(B\varphi_{11}, 0.90), (B\varphi_{12}, 0.75)$ $(B(\varphi_{11} \wedge \varphi_{12} \rightarrow \varphi_{17}), 1)$
$\mathcal{T}_D$	$(D\varphi_{17}, 0.95)$
$\mathcal{T}_I$	$(I\varphi_{17}, 0.95)$
$\mathcal{T}_N$	

TABLE 1. s_1's theories

φ_{11}	$\langle river\_basin, water\_index, adequate\rangle$
φ_{12}	$\langle rain\_fall, index, good\rangle$
φ_{17}	$\langle internal\_demand, status, satisfied\rangle$

TABLE 2: Propositions relevant for s_1's cognitions at t_1

In Table 1, we list the elements of the theories $\mathcal{T}_B, \mathcal{T}_D, \mathcal{T}_I$ and $\mathcal{T}_N$ of agent s before the new treaty was proposed where the propositions φ_{ij}'s are as given in TABLE 2. Agent s computes the coherence graphs with the nodes as the elements of TABLE 1, and their composite coherence graph $g_1 = \langle V_1, E_1, \zeta_1\rangle$. Then the coherence $\kappa(g_1)$

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

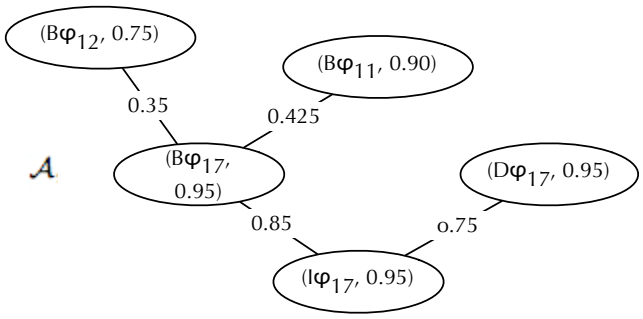

FIG. 6. Initial coherence graph (g_1) of s with $\kappa(g_1) = 0.59375$

is 0.59375. The accepted set $\mathcal{A} = V_1$ itself as the agent does not have any strong incoherence.

6.2.1 Evaluating the Treaty

Theory	*Existing*	*New*
$\mathcal{T}_N$		$(O\varphi_{13}, 1)$
$\mathcal{T}_B$	$(B\varphi_{11}, 0.90), (B\varphi_{12}, 0.75)$ $(B(\varphi_{11} \wedge \varphi_{12} \rightarrow \varphi_{17}, 1)$	$(B(\neg\varphi_{13} \rightarrow \varphi_{15}), 0.85)$ $(B\varphi_{11} \wedge \varphi_{12} \wedge \varphi_{13} \rightarrow \varphi_{17}), 0.9)$
$\mathcal{T}_D$	$(D\varphi_{17}, 0.95)$	$(D\neg\varphi_{15}, 1)$
$\mathcal{T}_I$	$(I\varphi_{17}, 0.95)$	$(I\neg\varphi_{15}, 1)$

TABLE 3. New elements into s's theories at t_1

φ_{13}	$\langle water\_release, quantity,\ 300\ billion\ ft^3\rangle$
φ_{15}	$\langle s_2\_threat, status, realised\rangle$
φ_{16}	$\langle treaty\_proposal, status, accepted\rangle$

TABLE 4. Propositions relevant for s_1's cognitions at t_1

Agent s evaluates the proposal of the new treaty by incorporating into its theories and its respective coherence graphs the new treaty, its implications and the sanctions that might be incurred if the proposal is not accepted. That is, the theories are updated as in TABLE 3, where the propositions φ_{ij}s are as in TABLE 4. Hence, the coherence graphs are updated with the new observations.

Agent s now computes the composite coherence graph $g_2 = \langle V_2, E_2, \zeta_2\rangle$ resulting from the theory update and using the set of bridge rules $\{b_1, b_2, b_3\}$. Using the bridge rules b_3 and b_4 as in Section 6.1, agent s reason about the new norm and its implications on its cognition. Since, the agent believes that $(B\varphi_{11} \wedge \varphi_{12} \wedge \varphi_{13} \rightarrow \varphi_{17}), 0.9)$, that is its intention of satisfying the internal demands can be met even while obeying the norm, we have positive coherence between $O\varphi_{13}, 1)$ and $(I\varphi_{17}, 0.95)$. Further

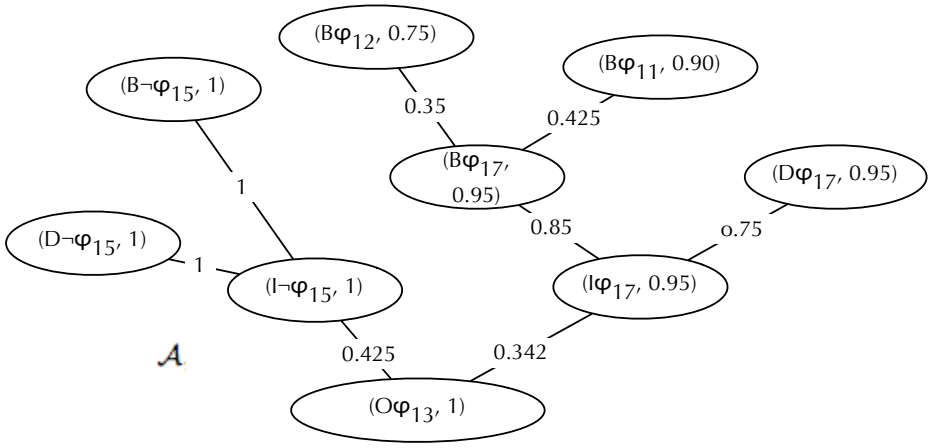

FIG. 7. Coherence graph (g_1), with norm accepted $\kappa(g_1) = 0.64275$

since s intent to avoid the sanction $(I\neg\varphi_{15}, 1)$, using the bridge rule b_4, it again has positive coherence between $O\varphi_{13}, 1)$ and $(I\neg\varphi_{15}, 1)$. Though the strength of each coherence relation is not very high, what we see is that $s$is in coherence with the new observation of the norm.

Applying the *max-cut* algorithm, we have the coherence maximising partition $(\mathcal{A}, V \setminus \mathcal{A})$ as shown in the Figure 7 with the accepted set $\mathcal{A}$ as V itself. The corresponding coherence of the graph, $\kappa(g_1)$ is 0.64275. It is clear that coherence has increased by incorporating the new norm. Hence, guided by coherence maximisation, agent s signs the treaty.

6.3 The Incoherence Buildup

Snapshot t_2 : 1991
Agent : s
New Observations: s experiences large-scale industrialisation, urbanisation, and higher revenue growth where as s also experiences higher water usage and a forecast of diminished rain fall.

To simplify the example, we take only a few representative observations from the above list as in TABLE 5 where the corresponding φ_{ij}s are as in TABLE 6.

The coherence graph g_3 of the agent s with changed cognitions is shown in Figure 8. Some of the coherence relations that do not influence the result have not been included in g_3 for the sake of clarity. Using the coherence equations, the coherence maximising partition $(\mathcal{A}, V \setminus \mathcal{A})$ is shown in Figure 8. The partition interestingly places the cognitions about φ_{22} and φ_{17} in set $\mathcal{A}$ while the cognitions about φ_{13} and $\neg\varphi_{15}$ in set $V \setminus \mathcal{A}$. It is clear from the coherence evaluation that all these intentions cannot coexist while maintaining the maximum coherence. That is, the agent has to choose between *obeying the norm, hence avoiding the threat of military action* or *satisfying the internal demands for water, hence economic progress.* Even though the ultimate decision can

Theory	*Existing*	*New*
$\mathcal{T}_N$	$(O\varphi_{13}, 1)$	
$\mathcal{T}_B$	$(B(\neg\varphi_{13} \rightarrow \varphi_{15}), 0.85)$ $(B\varphi_{21}, 0.90), (B\varphi_{22}, 0.85)$	$(B\varphi_{11}, 0.2), (B\varphi_{12}, 0.35), (B\varphi_{13}, 1)$ $(B(\varphi_{11} \wedge \varphi_{12} \rightarrow \varphi_{17}, 0.7)$ $(B(\varphi_{11} \wedge \varphi_{13} \rightarrow \neg\varphi_{17}), 0.90)$ $(B\varphi_{22} \rightarrow \varphi_{21}, 0.80)$ $(B\varphi_{17} \rightarrow \varphi_{22}, 0.75)$
$\mathcal{T}_D$	$(D\varphi_{17}, 0.95), (D\neg\varphi_{15}, 1)$	$(D\varphi_{22}, 0.85)$
$\mathcal{T}_I$	$(I\varphi_{17}, 0.95), (I\neg\varphi_{15}, 1)$	$(I\varphi_{13}, 1), (I\varphi_{22}, 0.85)$ $(I(\varphi_{17} \rightarrow \varphi_{22}), 0.75)$ $(I(\varphi_{13} \rightarrow \varphi_{17}), 0.05)$

TABLE 5. New elements into s's theories at t_2

φ_{21}	$\langle water\_usage, growth\_index, high\rangle$
φ_{22}	$\langle revenue, growth\_index, high\rangle$

TABLE 6. Propositions related to s_1's cognitions at snapshot t_2

vary from other considerations of the agent, a purely coherence maximising agent will choose to violate the norm in order to keep a maximal state of coherence. With this example we show how a coherence maximising agent evaluates norms in the context of its cognitions.

6.4 Discussion

Even though the example only demonstrates the case of a single norm, the same can be extended to cases where there are multiple norms and there is a need to choose among the norms. In terms of coherence, this is selecting a norm (or a set of norms) which maximises the coherence of the graph. By 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 are other criteria that need to be considered. Some of them already mentioned and represented in the graph are sanctions and rewards. Another important factor by which an agent makes a decision to adopt a norm is observing the behavior of other agents. We can represent this as cognitive models of other agents.

6.5 Computational Complexity

It has been shown that, it is possible to convert a coherence maximisation problem can be converted to an equivalent max-cut problem. As max-cut is an NP-complete problem it becomes clear that coherence maximisation is also an NP-complete problem. However neural network based algorithms give good approaximation. Thagard in his formalisation of coherence has given several implementations of coherence, with an extensive implementation of a neural network model called ECHO [29]. He also compares it with a max-cut implementation. We have extended Thagard's imple-

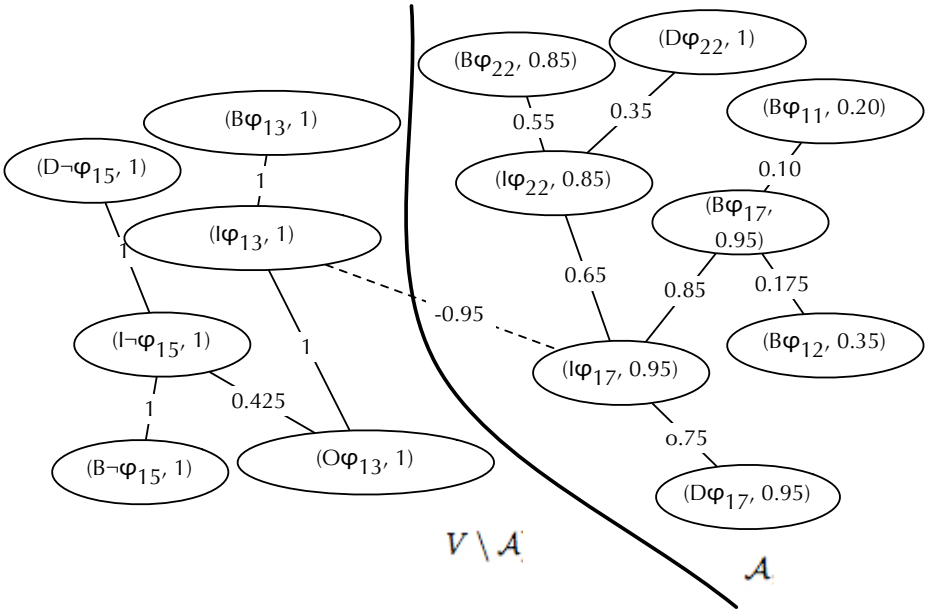

FIG. 8. Coherence graph (g_3), $\kappa(g_3) = 0.6769$

mentation to incorporate additional features of our model. That is, the solutions in our implementation we use a Prolog-based meta interpreter to extract proofs of each sentence in the BDI base of the agent where these proofs will give raise to the coherence values between pairs of sentences using the support function η of Section 3. We further use a semi-definite programming max-cut approximation algorithm to evaluate the coherence of the graph and to determine the nodes in the accepted set [31]. However, experimental evaluation of the case study is one 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. Due to the fact that our work links different areas of research, we in this section explore 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 filed 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 the agent design. The first major step was made when a behaviour model of agents was proposed. The BDI model for artificial agents is based on the theory of rational action in humans put forward in 1988 by the philosopher M. Bratman [6]. BDI is fundamentally reliant on folk psychology which is the notion that our mental models of the world are theories. BDI logics are multi-modal logics developed by Rao and Georgeff during the 1990s. However, the BDI model of agents was an attempt to solve a problem that has more to do with planning than with the design of autonomous agents. Yet, the BDI model served as the base model on which others could build more sophisticated features. From BDICTL of Rao and Georgeff's, LORA (the logic of rational agents) [33] to BOID [7], there have been numerous proposals to incorporate various levels of autonomy in agent design. However, as mentioned in the introduction, other than incorporating certain static priority based reasoning components into agent theories, we still lack sophisticated reasoning tools to make agents autonomous entities.

The work of Pasquier et al. [24] 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 which could be used to extend the agents reasoning mechanism to include social commitments. Their work is based, like ours, on the characterisation of coherence as maximising constraint satisfaction proposed by Thagard [29]. 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 which is fully computable and generic and in particular 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 [4]. One of the early introductions of norms for multiagent co-ordination is the work on artificial social systems by Tennenholtz and colleagues [28, 22, 16]. The problem studied in artificial social systems is the design, emergence or more generally the creation of social laws. Shoham and Tennen-

holtz 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 [10, 12, 34]. As our work mainly deals with the cognitive aspect of norms, the following discussion focuses on proposals from a cognitive perspective. We discuss two of the representative proposals below.

The work by Guido et al. [5] gives a comprehensive account of the situations faced by different types of agents in which they could possibly violate norms. Situations include: when there are contradictions between goals and obligations, when violation is preferred to possible sanction, when an agent is ignorant about a norm or consequences of it, or, when it is impossible to fulfil the obligation. This work also attempts to formalise some of these notions. What relates Guido et al.'s work and ours is that all these situations are somehow incoherent, and a coherence-driven agent can be used to model them. However their work does not address the reasoning within the agent.

The work of Conte et al. treats norms from the cognitive perspective of individual agents. They claim that some of the most important issues surrounding the study of norms are how agents can acquire norms, how agents can violate norms, and how an agent can be autonomous [12, 11]. In their work they address the issue of autonomous norm acceptance in agents and how that is instrumental to distributed norm formation and norm conformity in an agent society. The authors describe autonomous norm acceptance as a two step process, first recognising the norm issued by an external entity as a norm, and once the agent has accepted this norm, deciding to conform to it. The first step according to the authors would form the normative belief, and the second step would create the normative goal or intention. Moving from normative belief to normative conformity would additionally need the existence of other private goals of the agent which would benefit from the normative goal. The work provides a set of rules for normative acceptance and conformity. The authors, though recognising the importance of norm acceptance, sidestep the problem of coming up with mechanisms for autonomous norm acceptance. That is, recognising a norm as a norm is not equivalent to evaluating the norm. For an autonomous agent to accept a norm, the agent has to understand what a norm really means and its implications in terms of its own cognitions. And to conform to a norm it should know what actions or beliefs are permitted, prohibited or obliged.In this sense our work is complementary to theirs as we propose a mechanism for norm evaluation, which can be embedded in the process of norm emmergence proposed by the authors.

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

The work of Piwek in [25] attempts to model dialogue coherence in terms of gen-

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

The work of Valencia et al. [27] models agent dialogue based on the theory of dissonance. The theory of cognitive dissonance states that contradicting cognitions serve as a driving force that compels the mind to acquire or invent new thoughts or beliefs, or to modify existing beliefs, so as to reduce the amount of dissonance (conflict) between cognitions. Their work exploits the drive to reduce dissonance as a cause to initiate a dialogue and further when this dissonance no longer persists to terminate the dialogue. It is curious to note that many authors who have used the theory of dissonance in dialogue initiation and termination [24, 27] have not considered the fact that not all incoherences are dissonance. Further, dissonance seeks out specialised information or actions. The most important difference between the work of Valencia et al. and ours is that, for them coherence (or the lack of it) is a local phenomena concerning only the new arriving fact and the fact that it contradicts with, whereas for us coherence is a global phenomena affecting the entire knowledge base of the agent. As in the case of the previous work, the authors equate coherence with logical entailment.

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 [23, 2, 24]. Internal deliberation helps agents to deal with internal conflicts in goals, conflicts among goals and norms and so on. 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 pair-wise 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 that every argument

outside of it are defeated. These notions in a broad sense capture the idea of 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 coherence graph. First of all, a preferred extension attempts to find those sets of arguments that are un-disputable. They do not tolerate inconsistencies or contradictions. An argument based system tend to be more brittle in that they 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 also 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 with one another not in absolute values. One can often 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(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 [1], 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 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 models 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 [14] apply a coherence based mechanism for practical reasoning systems.

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 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 Future Work

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

However, there are criticisms raised about the philosophy of coherence. One question often put forward with respect to the application of coherence as an agent decision making tool is, whether it is rational for an agent to behave according to coherence maximisation. Normally an agent reasoning about norms takes into account influences of utility maximisation, models of other agents, and sanctions or rewards. We claim that we can introduce these decision making factors into our coherence graph so that the coherence maximisation is the only evaluation necessary for the decision making process. 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. However, as we have stated in Section 6.5, coherence maximisation is an NP-complete problem. However, as we assume bounded rationality for our agents, our coherence graphs are also bounded and are not complete. To reduce the computational burden further, our future work aims to bring in contexts, which would consider only a sub-graph of the actual with the intuition that coherence maximisation should consider only those nodes which are relevant to the problem at hand.

An important issue we have not explored in depth is the treatment of norms. We would like 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 is partially supported by the Open-

Knowledge[5] Specific Targeted Research Project (STREP), which is funded by the European Commission under contract number FP6-027253 and the Generalitat de Catalunya, under grant 2005-SGR-00093, and Spanish project "Agreement Technologies" (CONSOLIDER-INGENIO 2010 CSD2007-0022)

References

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

[2] Leila Amgoud, Nicolas Maudet, and Simon Parsons. Modeling dialogues using argumentation. In *ICMAS '00: Proceedings of the Fourth International Conference on MultiAgent Systems*. IEEE Computer Society, 2000.

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

[4] Guido Boella, Leendert Torre, and Harko Verhagen. Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, 12(2-3), 2006.

[5] 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*, 2004.

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

[7] Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *AGENTS '01: Proceedings of the fifth international conference on Autonomous agents*, pages 9–16. ACM, 2001.

[8] Ana Casali, Llus Godo, and Carles Sierra. Graded BDI models for agent architectures. In *Lecture Notes in Computer Science*, volume 3487, 2005.

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

[10] Cristiano Castelfranchi, Frank Dignum, Catholijn M. Jonker, and Jan Treur. Deliberative normative agents: principles and architecture. In *ATAL '99: 6th International Workshop on Intelligent Agents VI, Agent Theories, Architectures, and Languages*, pages 364–378. Springer-Verlag, 2000.

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

[12] Rosaria Conte, Cristiano Castelfranchi, and Frank Dignum. Autonomous norm acceptance. In *The Sixth International Workshop on Agent Theories, Architectures, and Languages*. Springer-Verlag, 1998.

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

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

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

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

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

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

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

[20] Martin J. Kollingbaum and Timothy J. Norman. Norm adoption in the noa agent architecture. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 1038–1039, 2003.

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

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

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

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

[24] Philippe Pasquier and Brahim Chaib-draa. The cognitive coherence approach for agent communication pragmatics. In *Second International Joint Conference on Autonomous Agents and Multiagent Systems*, 2003.

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

[26] Anand S. Rao and Michael P. Georgeff. Bdi agents: From theory to practice. In *ICMAS–95, First International Conference on Multi-Agent Systems: Proceedings*, pages 312–319. MIT Press, 1995.

[27] Jean-Paul Sansonnet and Erika Valencia. A model for dialog between semantically heterogeneous informational agents. In *Eleventh Portuguese Conference on Artificial Intelligence*, 2003.

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

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

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

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

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

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

[34] Fabiola López y López, Michael Luck, and Mark d'Inverno. Constraining autonomy through norms. In *First International Joint Conference on Autonomous Agents and Multiagent Systems*, 2002.

Received 8 December 2008.