A Coherence Based Framework for Institutional Agents

Sindhu Joseph, Carles Sierra, and Marco Schorlemmer

Artificial Intelligence Research Institute, IIIA Spanish National Research Council, CSIC
Bellaterra (Barcelona), Catalonia, Spain
{joseph,sierra,marco}@iiia.csic.es

Abstract. We introduce in this paper an agent model based on coherence theory. We give a formalization of Thagard's theory on coherence and use it to explain the reasoning process of an intentional agent that permits the agent to drop beliefs or to violate norms in order to keep a maximal state of coherence. The architecture is illustrated in the paper and a discussion on the possible use of this approach in the design of institutional agents is presented.

1 Introduction

Artificial institutions are multiagent system models inspired by human institutions [10] and used to create technological extensions of human societies [12]. These devices are designed to help agents cope with the uncertainty on the environment and in some cases to increase their individual utility. They are important due to the bounded nature of human and software rationality (global maximization of individual utility cannot be guaranteed in a complex society). If two or more persons exchange goods with one another, then the result for each one will depend in general not merely upon his own actions but on those of the others as well [8]. Therefore, to make these exchanges possible, behavioral rules that govern the way in which individuals can cooperate and compete are required [7]. Behavioral rules translate the social objectives into executable permissions, prohibitions, and obligations. These modalities are collectively called *norms*. Thus, institutions are role based normative systems representing a collective intention[1]. This is the case in general, but we do acknowledge the fact that institutions need not always represent a collective intention. But such institutions almost always undergo periodic revolutions as an attempt to reinforce collective intention.

Human institutions tend to adapt when the group conscience shifts or is in conflict with the current institutional definition. It is thus important to know and be able to verify at any point in time, that the institutional definition in *coherence* with its norms and social objectives and the objectives of the individuals in the group. Thus an institution to be sustainable almost always needs to continuously strive to achieve this coherent state, here we call it *equilibrium*. We say an institution is in a state of equilibrium when it has no incentive to change the institutional definition. When an *incoherence* or a deviation from equilibrium is detected, it is also important to identify the candidates that cause this incoherence to be able to bring the institution back into equilibrium.

[1] Collective intention here refers to the explicit expression of the intention and do not refer to the mental state.

An autonomous agent is motivated to join an institution when it believes that the individual goals of the agent can be satisfied within the institution. And that happens in our opinion when the beliefs or goals of the agent are *coherent* with the institutional objectives. For simplicity, here we assume that all institutional objectives are realized through norms. Thus being incoherent with a norm is equivalent to being incoherent with a corresponding institutional objective. An agent will hence need to continuously re-evaluate the alignment of its beliefs and goals with that of the norms of the institution. Thus, it is important for an agent to know whether there is an incoherence among the beliefs and the norms, and how the decision is made on what needs to be changed to bring the coherence back. This incoherence among other things drives the agent to violate a norm, revise a belief or both. The individual state of equilibrium is achieved when the coherence between individual beliefs and goals, those of the group and those of the institution is maximized.

We use the theory of coherence and the theory of cognitive dissonance to ground our framework. The *theory of coherence* [11] has been well studied in the field of cognitive science and as a general theory to describe the world. Coherence theory is about how different pieces fit together to make a whole. It assumes that there are various kinds of associations between the pieces or the elements of a set. These are primarily positive or negative where a positive association suggests that the two elements support each other while a negative association indicates their mutual exclusion. Thagard views these associations as constraints between elements and proposes a theory of coherence as globally maximizing the satisfaction of these constraints. He proposes to partition the set of elements into accepted or rejected so that the overall coherence is achieved, or constraint satisfaction maximized. We use the theory to reason between the cognitions of an agent and its external associations such as institutions or social relations.

The *theory of dissonance* [5] in social psychology is closely related to the theory of coherence. Leon Festinger calls dissonance as the distressing mental state in which people feel they "find themselves doing things that don't fit with what they know, or having opinions that do not fit with other opinions they hold." The tension of dissonance motivates us to change either our behavior or our belief in an effort to avoid a distressing feeling. The more important the issue and the greater the discrepancy between behavior and belief, the higher the magnitude of dissonance that we will feel. We use the dissonance theory to motivate an action once the coherence theory identifies elements causing a reduction in coherence.

In this paper we propose an institutional agent architecture based on the theory of coherence. This architecture permits us to talk about the coherence of the individual beliefs, desires and intentions[2], coherence among these cognitions, and the coherence among the cognitions and institutional norms or social commitments. In particular when there is an incoherence between any of these elements, the agent often needs to choose between a norm violation or a belief revision to maximize its internal coherence. That is, the theory of incoherence helps us to model autonomous agents who can reason about obeying or violating institutional norms. From the institutional point of view, the same tools can be used to reason about an institution, coherence of an institution with respect to the conscience of the group and how to evolve norms to stay in alignment

[2] In the paper we discuss beliefs, the extension to desires and intentions is straight-forward.

with the objectives. While coherence theory helps to find the maximally coherent state, dissonance theory helps to decide how much of incoherence an agent or an institution can tolerate and which of the actions to chose from to reduce incoherence.

In Sections 2 and 3 we introduce our coherence-based framework and the reasoning of a coherence-maximizing agent. In Section 4 we illustrate with the help of an example, how this framework can be used to reason about norm violations. We conclude with related work in Section 5 and discussion and future work in Section 6. We use the example of a car agent in a traffic control institution. Here we give an intuitive summary of the example, for the reader to follow the coherence framework introduced in Section 2. In Section 4, we detail the example further.

The car agent in our example has personal beliefs and intentions. Where-as the traffic control institution has a set of objectives which it implements through a number of norms. The car agent initially starts with the belief that the traffic control is efficient, and is in a maximally coherent state with his beliefs, intentions and institutional norms. But when the car agent reaches a crossing of two lanes and is made to stop at the signal, where as the crossing lane has no cars waiting to go, it builds up a certain incoherence with its other beliefs and intentions such as the intention to reach the destination in time and the belief that the traffic control is efficient. As part of the constraint maximization, the agent identifies that the adopted intention *to obey the traffic norms* should be rejected to restore coherence. Further it finds that the dissonance is high enough to actually reject this intention. After the rejection of the intention means a potential norm violation as it no longer considers to obey the traffic norms.

2 Coherence framework

In this section we introduce a number of definitions to build the coherence framework. Our primary interest is to put the theory in relation to an institutional agent context and to provide a formal representation and some computing tools. We do this for the belief cognition of an agent and for the norms of an institution.

2.1 Coherence Graph

To determine the coherence of a set of elements, we need to explore their associations. We shall use a graph to model these associations in order to compute coherence of various partitions of a given set of elements, and to determine its maximally coherent partition as well as to study other related aspects of coherency.

We shall define a coherence graph over an underlying logic. Given a set of propositional formulae PL, a *logic* over PL is a tuple $\mathcal{K} = \langle \mathcal{L}, A, \vdash \rangle$, with language $\mathcal{L} \subseteq PL \times [0, 1]$, i.e., a set of pairs formed by a proposition and a confidence value between 0 and 1, a set of *axioms* $A \subseteq \mathcal{L}$, and a *consequence relation* $\vdash \subseteq 2^{\mathcal{L}} \times \mathcal{L}$.

The nodes of a coherence graph are always elements of $\mathcal{L}$. The consequence relation determines the relationship between these elements, and thus puts constraints on the edges of a coherence graph. Furthermore, propositions that are assumed to be true belong to the axioms A of the logic.

A coherence graph is therefore a set ($\in V$) of nodes taken from $\mathcal{L}$ and a set E of edges connecting them. The edges are associated with a number called the *strength of the connection* which gives an estimate of how coherent the two elements are[3]. The strength value of an edge (φ, γ), noted $\sigma(\varphi, \gamma)$, respects the strength values that it has with other connected edges. It is important to note that a coherence graph is a *fully connected graph* with a restriction that for every node φ[4] $\in \mathcal{L}$, $\sigma(\varphi, \varphi) = 1$ and if there are two nodes φ and, ψ that are not related, then $\sigma(\varphi, \psi) = 0$. Further α is a projection function defined from the set V to $[0, 1]$ which projects the confidence degrees associated with elements of $\mathcal{L}$. The role of this function is to make the confidence degrees explicit in the graph for ease of explanation.

Definition 1. *Given a logic $\mathcal{K} = \langle \mathcal{L}, A, \vdash \rangle$ over a propositional language PL, a* coherence graph $\langle V, E, \sigma, \alpha \rangle$ *over $\mathcal{K}$ is a graph for which*

- $V \subseteq \mathcal{L}$
- $E = V \times V$
- $\sigma : E \rightarrow [-1, 1]$
- $\alpha : V \rightarrow [0, 1]$

and which satisfies the following constraints:

- $A \subseteq V$
- $\forall v \in V, \sigma(v, v) = 1$
- $\sigma(v, w) = \sigma(w, v)$

We write $\mathcal{G}(\mathcal{K})$ for the set of all coherence graphs over $\mathcal{K}$.

Given this general definition of a *coherence graph*, we can instantiate two specific families of coherence graphs namely the *belief coherence graphs* $\mathcal{BG}$ and the *norm coherence graphs* $\mathcal{NG}$, which are of interest to us. $\mathcal{BG}$ represents graphs where the nodes are beliefs of an agent and the edges are association between beliefs. And $\mathcal{NG}$ represents nodes which are the possible norms defined in an institution. In this paper, we do not discuss the desire and the intention cognitions, but these can be defined similarly. And when defining the norm logic, we only talk about permissions and obligations, whereas norms may include prohibitions, too. Also for clarity we have kept the structure of the norms simple, but we intend to include objectives and values associated with a norm. The work by Atkinson and Bench-Capon [1] is indicative. We now define the belief and the norm logic to express the nodes of these graphs and their interconnections.

In our representation, beliefs are propositional formulas φ which are closed under negation and union with an associated confidence degree d. We may borrow the axioms and the consequence relation $\vdash$ from an appropriate belief logic. Then for example we have the following definition for the belief logic.

[3] This value is fuzzy and is determined by the type of relation between the edges. For an *incoherence* relation, tends toward -1, for *coherence* a positive value tending toward 1.

[4] This should be understood as $\langle \varphi, d \rangle$, whenever it is understood from the context, we omit the d part of the element for better readability.

Definition 2. *Given the propositional language PL, we define the* belief logic $\mathcal{K}_B = \langle \mathcal{L}_B, A_B, \vdash_B \rangle$ *where*

- *the belief language $\mathcal{L}_B$ is defined as follows:*
  - *Given* $\varphi \in PL$ and $d \in [0,1]$, $\langle B\varphi, d\rangle \in \mathcal{L}_B$
  - *Given* $\langle \theta, d\rangle, \langle \psi, e\rangle \in \mathcal{L}_B$, $\langle \neg\theta, f(d)\rangle \in \mathcal{L}_B$ and $\langle \theta \wedge \psi, g(d,e)\rangle \in \mathcal{L}_B$ where f and g are functions for example as in [3]
- A_B *as axioms of an appropriate belief logic.*
- $\vdash_B$ *is a consequence relation of an appropriate belief logic.*

We need a number of additional constraints that we want the Belief coherence graphs to satisfy. They are constraints on how the strength values have to be assigned. A constraint that we impose on this number is that if two elements are related by a $\vdash$, then the value should be positive and if two elements contradicts then then there is a negative strength[5]. And here we define α more concretely as the projection function over the belief degree. Then we have

Given the belief logic $\mathcal{K}_B$, the *set of all belief coherence Graphs is* $\mathcal{G}(\mathcal{K}_\mathcal{B})$ satisfying the additional constraints:

- Given $\varphi, \psi \in \mathrm{V}$ *and* $\Gamma \subseteq \mathrm{V}$ *and* $\Gamma \vdash \varphi$
  - $\forall \gamma \in \Gamma, \sigma(\varphi, \gamma) > 0$
  - $\forall \gamma \in \Gamma$ *and* $\psi = \neg\varphi, \sigma(\psi, \gamma) < 0$
- $\forall \langle B\varphi, d\rangle \in V, \alpha(\langle B\varphi, d\rangle) = d$

We can similarly derive the set of all norm coherence graphs $\mathcal{G}(\mathcal{K}_N)$ corresponding to norms. In our definition, norms define obligations and permissions associated with a role. We use *deontic logic* to represent the norms, with the difference that we use modalities subscripted with roles. Thus O_r and P_r represent deontic obligations and deontic permissions associated with a role $r \in R$, the set of all roles. In this paper we assume the confidence degrees associated with norms to be 1. Thus we have the following definition for a norm logic $\mathcal{K}_N$.

Definition 3. *Given the propositional language PL and the set of roles$R$, we define the* Norm logic $\mathcal{K}_N = \langle \mathcal{L}_N, A_N, \vdash_N \rangle$ *where*

- $\mathcal{L}_N$ *is defined as:*
  - Given $\varphi \in PL$ *and* $r \in R$, *then* $\langle O_r\varphi, 1\rangle, \langle P_r\varphi, 1\rangle \in \mathcal{L}_N$
  - Given $\langle \varphi, d\rangle$ and $\langle \psi, e\rangle \in \mathcal{L}_N$ then $\langle \neg\varphi, f_1(d)\rangle$ and $\langle \varphi \wedge \psi, g_1(d,e)\rangle \in \mathcal{L}_N$
- A_N *following the standard axioms of deontic logic.*
- $\vdash_N$ *using the standard deduction of deontic logic*[6]

Given the norm logic $\mathcal{K}_N$ *the set of all norm coherence graphs* is $\mathcal{G}(\mathcal{K}_N)$ satisfying the additional constraints:

- Given $\varphi, \psi \in \mathcal{L}$ *and* $\Gamma \subseteq \mathcal{L}$ *and* $\Gamma \vdash \varphi$
  - $\forall \gamma \in \Gamma, \sigma(\varphi, \gamma) > 0$
  - $\forall \gamma \in \Gamma$ *and* $\psi = \neg\varphi, \sigma(\psi, \gamma) < 0$
- $\forall \langle \varphi, d\rangle \in V, \alpha(\langle \varphi, d\rangle) = 1$

[5] This relates to Thagard's *deductive coherence*, though in this paper, we limit our discussion to the general coherence relation.

[6] For an introduction to deontic logic, see [13] and in the context of institutions see [6]

2.2 Calculating Coherence

We can now define the coherence value of a graph, the partition that maximizes coherence and the coherence of an element with respect to the graph. These values will help an agent to determine whether to keep a belief or drop it, whether to obey a norm or violate it to increase coherence and which of the beliefs or norms need to be dropped to maximize coherence. This will also help an institution decide whether to accept a proposed norm change and to determine the gain in coherence when accepting or rejecting a change.

We use the notion of coherence as maximizing constraint satisfaction as defined by Thagard [11]. The intuition behind this idea is that there are various degrees of coherence/incoherence relations between nodes of a coherence graph. And if there is a strong negative association between two nodes, then the graph will be more coherent if we decide to accept one of the nodes and reject the other. Similarly when there is a strong positive association, coherence will be increased when either both the nodes are accepted or both are rejected. Thus we can construct a partition of the set of nodes, with one set of nodes in the partition being accepted and the other rejected in such a way to maximize the coherence of the entire graph. Such accepted sets are denoted by $\mathcal{A}$ and the rejected sets by $\mathcal{R}$. The coherence value is calculated by considering positive associations within nodes of $\mathcal{A}$ and within nodes of $\mathcal{R}$ and negative associations between nodes of $\mathcal{A}$ and $\mathcal{R}$. This criteria is called *satisfaction of constraints*. More formally we have the following definition:

Definition 4. *Given a coherence graph $g \in \mathcal{G}(\mathcal{K})$ and a partition $(\mathcal{A}, \mathcal{R})$ of V, we define the* set of satisfied associations $C^+ \subseteq E$ *as*

$$C^+ = \left\{ \forall (v_i, v_j) \in E \;\middle|\; \begin{array}{l} v_j \in \mathcal{A} \leftrightarrow v_i \in \mathcal{A} (\text{or } v_j \in \mathcal{R} \leftrightarrow v_i \in \mathcal{R}) \text{ when } \sigma(v_i, v_j) \geq 0 \\ v_j \in \mathcal{A} \leftrightarrow v_i \in \mathcal{R} \text{ when } \sigma(v_i, v_j) < 0 \end{array} \right\}$$

In all other cases the association is said to be unsatisfied.

To define coherence, we first define the total strength of a partition. The total strength of a partition is the sum of the strengths of all the satisfied constraints multiplied by the degrees (the α values) of the nodes connected by the edge. Then the coherence of a graph is defined to be the maximum among the total strengths when calculated over all its partitions. We have the following definitions:

Definition 5. *Given a coherence graph $g \in \mathcal{G}(\mathcal{K})$, we define* the total strength of a partition $\{\mathcal{A}, \mathcal{R}\}$ *as*

$$S(g, \mathcal{A}, \mathcal{R}) = \sum_{(v_i, v_j) \in C^+} \mid \sigma(v_i, v_j) \mid \cdot \alpha(v_i) \cdot \alpha(v_j) \tag{1}$$

Definition 6. *Given a coherence graph $g = \langle V, E, \sigma, \alpha \rangle \in \mathcal{G}(\mathcal{K})$ and given the total strength $S(g, \mathcal{A}, \mathcal{R})$ for all partitions of V (denoted as $\mathcal{P}(V)$), we define* the coherence of g *as*

$$C(g) = \max\{S(g, \mathcal{A}, \mathcal{R}) \mid \mathcal{A}, \mathcal{R} \in \mathcal{P}(V)\} \tag{2}$$

and we say that the partition with the maximal value divides the set of nodes into an accepted set $\mathcal{A}$ and a rejected set $\mathcal{R}$.

Given the coherence $C(g)$ of a graph, *the coherence of an element* $C(\varphi)$ is the ratio of coherence when φ is in the accepted set with respect to φ not being in the accepted set. That is if the acceptance of the element improves the overall coherence of the set considered, than when it is rejected, then the element is said to be coherent with the set. Then we have the definition:

Definition 7. *Given a coherence graph $g \in \mathcal{G}(\mathcal{K})$, we define the* coherence of an element $\varphi \in V$ *as*

$$C(\varphi) = \frac{\max_{\substack{\mathcal{A},\mathcal{R}\in\mathcal{P}(V)\\ \varphi\in\mathcal{A}}} S(g,\mathcal{A},\mathcal{R})}{\max_{\substack{\mathcal{A},\mathcal{R}\in\mathcal{P}(V)\\ \varphi\notin\mathcal{A}}} S(g,\mathcal{A},\mathcal{R})} \tag{3}$$

Similar to the coherence definitions of a graph, we now define the dissonance of a graph. We define dissonance as the measure of incoherence that exists in the graph. Deducing from the theory of dissonance [5] an increase in dissonance increases in an agent the need to take a coherence maximizing action. We use the dissonance as a criteria to chose among the number of alternative actions an agent can perform such as belief revision, norm violation or commitment modification for example. The dissonance of a graph is computed as the difference between the total strength of the graph and the coherence of the graph. Thus we have the following definition:

Definition 8. *Given a coherence graph $g \in \mathcal{G}(\mathcal{K})$, we define the dissonance of g with respect to a partition $(\mathcal{A},\mathcal{R})$ as*

$$D(G,\mathcal{A},\mathcal{R})^{7} = \begin{cases} \infty \text{ if } C(G) = 0 \\ \frac{C(G)-S(G,\mathcal{A},\mathcal{R})}{C(G)} \text{ otherwise} \end{cases} \tag{4}$$

2.3 Graph Composition

For an agent that is part of an institution and has social relations, it not only needs to maximize the internal coherence between its beliefs, but also needs to maximize the *social coherence* which is the coherence between the beliefs and the commitments made in the context of his social relations. Similarly, an agent which belongs to an institution, needs to maximize the *institutional role coherence*, that is the coherence between the projection of the norms onto the role he plays in the institution and his beliefs. This leads naturally the notion of graph composition, which will allow us to explore the coherence or incoherence that might exist between nodes of one graph and those of the other.

The nodes of a composite graph are always the disjoint union of the nodes of the individual graphs. The set of edges contains at least those edges that existed in the individual graphs. In addition a composite graph may have new edges between nodes of one graph to the nodes of the other graph.

Definition 9. *Let $\mathcal{K}_1 = \langle \mathcal{L}_1, A_1, \vdash_1 \rangle$ and $\mathcal{K}_2 = \langle \mathcal{L}_2, A_2, \vdash_2 \rangle$ be logics over propositional language PL_1 and PL_2. Let $g_1 = \langle V_1, E_1, \sigma_1, \alpha_1 \rangle \in \mathcal{G}(\mathcal{K}_1)$ and $g_2 =$*

[7] When $C(G) = 0$, $S(G,\mathcal{A},\mathcal{R}) = 0$ and hence the dissonance is maximum. $D(G,\mathcal{A},\mathcal{R}) = \infty$

$\langle V_2, E_2, \sigma_2, \alpha_2\rangle \in \mathcal{G}(\mathcal{K}_2)$. *The* set of composite graphs $g_1 \odot g_2 \subset \mathcal{G}(\mathcal{K})$ *is the set of those coherence graphs* $\langle V, E, \sigma, \alpha\rangle \in \mathcal{G}(\mathcal{K})$ *over logic* $\mathcal{K} = \langle \mathcal{L}, A, \vdash\rangle$*—where* $\mathcal{L}$ *is the disjoint union of* $\mathcal{L}_1$ *and* $\mathcal{L}_2$*,* A *is the disjoint union of* A_1 *and* A_2*, and* $\vdash$ *is the smallest consequence relation containing both* $\vdash_1$ *and* $\vdash_2$[8]*— such that*

- $\mathrm{V} = \{\mathcal{L}_1/\varphi \mid \varphi \in \mathrm{V}_1\} \cup \{\mathcal{L}_2/\varphi \mid \varphi \in \mathrm{V}_2\}$[9]
- $\mathrm{E} = \mathrm{V} \times \mathrm{V}$ such that
  - if$(\varphi, \psi) \in \mathrm{E}_1$ then $(\mathcal{L}_1/\varphi, \mathcal{L}_1/\psi) \in \mathrm{E}$
  - if$(\varphi, \psi) \in \mathrm{E}_2$ then $(\mathcal{L}_2/\varphi, \mathcal{L}_2/\psi) \in \mathrm{E}$
- $\sigma : \mathrm{E} \to [-1, 1]$ such that
  - $\sigma(\mathcal{L}_1/\varphi, \mathcal{L}_1/\gamma) = \sigma_1(\varphi, \gamma)$
  - $\sigma(\mathcal{L}_2/\varphi, \mathcal{L}_2/\gamma) = \sigma_2(\varphi, \gamma)$

These properties state that the nodes of the composite graph are the disjoint union of the original graphs. When making the composition, the existing edges and strength values are preserved.

3 A coherence maximizing agent

In this section we describe some of the reasoning performed by a coherence maximizing agent. Consider an agent a having a belief coherence graph b, intention coherence graph i and role coherence graph n_r. At any moment in time the agent aims at coherence maximization. When the coherence cannot be further maximized, a does nothing, or has no incentive to act. For an agent who has no social commitments, nor is part of any institution, nor has any unfulfilled intentions, the accepted set $\mathcal{A}$ is the entire belief set, as he is not likely to have an incoherence.

We consider an agent that is part of an institution, has social commitments and is in the state of equilibrium. Below we show one of the possible algorithms that a coherence agent a can go through when it encounters a new belief (either communicated to the agent by others, by observation, or internally deduced).

Input: a new belief $\langle B\varphi, d\rangle$; a belief coherence graph $g = \langle V, E, \sigma, \alpha\rangle$, a composition graph $g_{bin} = g \odot g_i \odot g_{n_r}$ with the corresponding coherence measures C_{bin} along with $\mathcal{A}_{bin}$ and $\mathcal{R}_{bin}$, S_{bin}, D_{bin}, and a dissonance threshold D_T.

1: $V_b \leftarrow V \cup \{B\varphi\}$
2: $\alpha_b(B\varphi) \leftarrow d$
3: **for** $B\psi \in V$, $\Gamma \subseteq V$ **do**
4: **if** $B\psi, \Gamma \vdash B\varphi$ or $B\varphi, \Gamma \vdash B\psi$ **then**
5: $\sigma_b(B\psi, B\varphi) = 1$
6: **for** $B\gamma \in \Gamma$ **do**
7: $\sigma_b(B\gamma, B\varphi) = 1$
8: **end for**
9: **end if**

[8] For the moment we assume that the properties that make $\vdash_1$ and $\vdash_2$ a consequence relation as the same.

[9] We write $\mathcal{L}_i/\varphi$ for those elements of $\mathcal{L}$ that come form $\mathcal{L}_i$ in the disjoint union, with $i = 1, 2$.

10: **if** $B\varphi, B\psi \vdash \perp$ **then**
11: $\sigma_b(B\varphi, B\psi) = -1$
12: **end if**
13: **end for**
14: $g_{bin} \leftarrow g_b \odot g_i \odot g_{n_r}$
15: $S \leftarrow S_{bin}(g_{bin}, V_{bin}, \emptyset)$ using eq(1)
16: $C \leftarrow C_{bin}(g_{bin})$ using eq(2)
17: $D \leftarrow D_{bin}(g_{bin}, V_{bin}, \emptyset)$ using eq(4)
18: **if** $D \geq D_T$ **then**
19: $\mathcal{A} \leftarrow \mathcal{A}_{bin}$
20: $\mathcal{R} \leftarrow \mathcal{R}_{bin}$
21: **end if**

The lines from 1 to 13 updates the belief graph by adding nodes, edges and their strength values. Here the algorithm does not fully determine the strength values but specify certain constraints on how the strength values are determined. Here we assume that a human user will provide them while respecting the constraints though we envision many semi automatic methods worth exploring (see section 6). The line 14 updates the composition graph considering the modified belief graph. The lines from 15 to 17 recalculate the strength, coherence and dissonance values of the new composite graph. Lines 18 and 19 check whether the dissonance value exceeds the threshold and if it does, the agent acts by removing the nodes causing the incoherence from the accepted set. To keep the discussion simple in this algorithm, we have simply removed the nodes. But in reality, the reaction to an incoherence can vary greatly. For instance a mildly distressed agent may choose to ignore the incoherence, may be satisfied with lowering the degree associated with a particular belief, may still choose to follow a norm. Where as a heavily distressed agent may not only chose to violate a norm, but initiate a dialogue to campaign for a norm change.

4 An Example

The main entities in our example are a car agent a having the role c in a traffic control institution and the institution itself T. We take a very simplified version of the objectives of T as

- minimizing the probability of collisions
- increasing the traffic handling capacity

To meet these objectives, the traffic control system has a signal at the crossing of the lanes along with specific norms of use. The norms of the traffic control system for the car agents belong to the set $\mathrm{N_c}$.

The traffic is controlled using the norms given below and the corresponding norm coherence graph is shown in Figure 1. Note that all the coherence graphs in this example have additional self loops which are not drawn for the sake of readability. But it is included in the coherence calculations.

- $\mathrm{O_c}(\mathrm{RED} \rightarrow \mathrm{STOP}), 1 \rightarrow$ *It is obligatory to STOP, when the signal is RED*
- $\mathrm{P_c}(\mathrm{GREEN} \rightarrow \mathrm{GO}), 1 \rightarrow$ *It is permitted to GO , when the signal is GREEN*

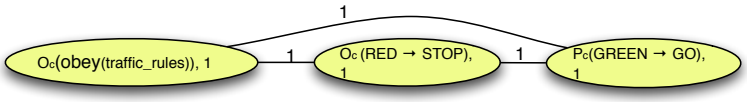

Fig. 1. Norm Coherence graph of the traffic control institution

Here we illustrate the model with one of the most simple cases, namely the crossing between a major and a minor lane. The major lane has more traffic than minor lane. Due to the fixed time control, and due to ignoring to assign priority to the lanes, the signal durations are the same for both major and minor lanes. Thus there are situations when there are no cars waiting to cross at the minor lane and there is a "RED" light at the major lane. So the car agents at the major lane sometimes experience an incoherence when trying to follow the traffic norms. We now show the evolution of the coherence of an agent situated at the major lane with the help of the some figures.

A car agent a of role c at the major lane has the intention to reach destination X at time T. He holds a number of beliefs which support this intention. A few relevant beliefs of a for this intention are *can reach destination X in time t* and *traffic control is efficient* and a generic belief that *It is good to reduce pollution*. The composite graph $b \odot i$ is shown in Figure 2.

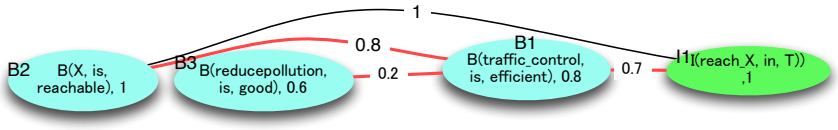

Fig. 2. $b \odot i$ Coherence graph of the car agent

We use Equations 1, 2, 4 of Section 2 for calculating the various coherence values of all the graphs of the example[10].

The coherence of the graph is $C(b \odot i) = 5.296$ with $\mathcal{A} = \{B1, B2, B3, I1\}$ and $D(b \odot i) = 0$. As a is part of the traffic control system, having a role c, the projection of the norms n_c to the beliefs graph of a with an additional intention *to stop at RED signal* is as given in Figure 3. This additional intention is due to the fact that a intends to follow the norms of the institution. Now the coherence of the composite graph is $C(b \odot i \odot n_c) = 17.716$ with $\mathcal{A} = \{B1, B2, B3, I1, I2, N1, N2\}$ and dissonance $D(b \odot i \odot n_c) = 0$, still staying 0.

When a encounters the "RED" signal, and observes the traffic, its belief graph gets enriched with new information, and due to this addition of new beliefs, the strengths get

[10] The strength values and the degrees on beliefs and intentions are given manually respecting the constraints on the graph definition.

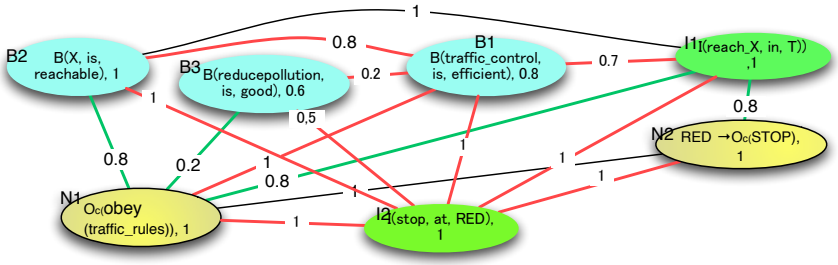

Fig. 3. Belief Coherence graph of the car agent with projected norms

modified. The new beliefs added to b are a *is at the Major lane*, *The signal is "RED"* and that *there are no cars on the minor lane*. The modified coherence graph is shown in Figure 4.

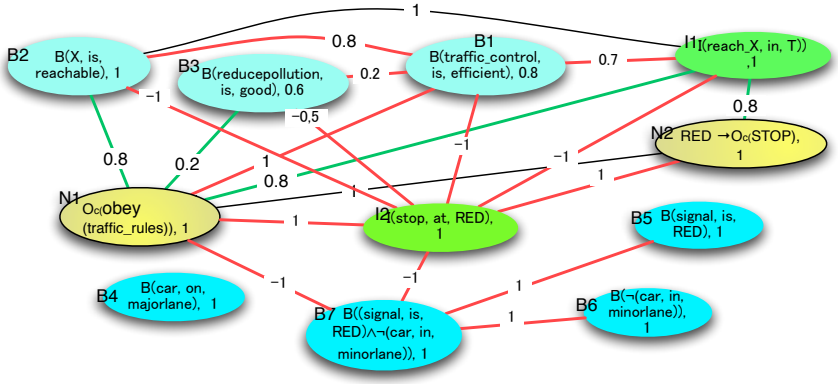

Fig. 4. Modified coherence graph

Now when trying to maximize the coherence, a discovers that if it removes the intention $I2 \rightarrow$ *to stop at RED signal* from the accepted set, he is able to maximize the coherence as in Figure 5. The total strength is $S(b \odot i \odot n_r, V, \emptyset) = 15.516$, Coherence of the graph is $C(b \odot i \odot n_r) = 23.716$ with $\mathcal{A} = \{B1, B2, B3, B4, B5, B6, B7, I1, N1, N2\}$ and dissonance $D(b \odot i \odot n_r) = 0.35$. Here the agent has a high enough dissonance[11] to reject the intention $I2$(*intention to obey the traffic norms*. This example though simple, illustrates how an agent can act based on coherence maximization.

[11] Assuming a dissonance threshold $D_T = 0.20$.

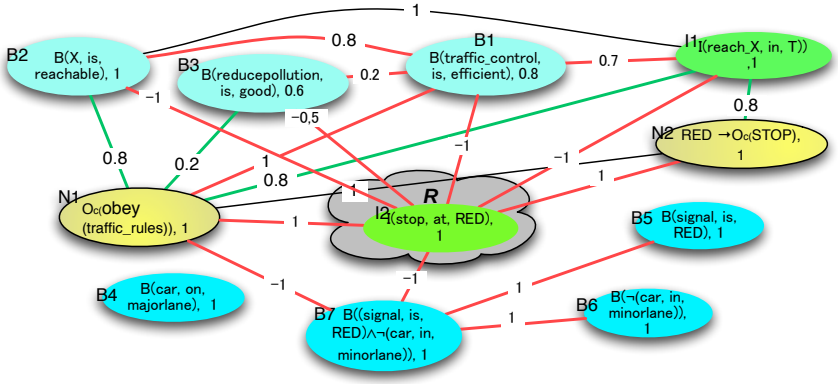

Fig. 5. maximizing coherence - $\mathcal{A} = b \odot i \odot n \setminus \{I2\}$

5 Related work

BDI theory is the most popular of the existing agent architectures. This architecture concentrates on the deliberative nature of the agent. There are several add ons to BDI architecture considering the recent developments in social and institutional agency, where the traditional cognitive model seems inadequate. They primarily include the addition of $norms$ to the cognitive concepts of *belief, desire, and intention*. The BOID architecture with the addition of obligation [2], and the work on deliberative normative agents [4] are the most prominent among them. In the BOID architecture the main problem is conflict resolution between and within the modules belief, desire, intention and obligation. Their focus is on architecture, while they do not specify any means to identify or resolve conflicts arising from interactions of B, O, I and D. Further the modules are flat structured where the associations between elements in the modules are not exposed making it difficult to identify and analyze conflicts. The work by Castelfranchi in [4] again concentrates on the architecture. Their main contribution is the emphasis on agent autonomy. While most literature assume the strict adherence to the norms, they insist that it is an agent's decision whether to obey norms or not. As in the BOID architecture, they do not provide any mechanism by which an agent can violate a norm or reason about a norm violation. Another work by Lopez et al. [14] discusses how norm compliance can be ensured while allowing autonomy, using rewards and sanctions. Such mechanisms, while certainly complimenting our approach, only handle the issue at a superficial level and do not give the power to an agent to understand what it means to obey or violate a norm with respect to its cognitions.

On the other hand, the work of Pasquier et al [9] is the first to our knowledge that attempts to unify the theory of coherence with the BDI architecture. The authors propose the theory as a reasoning mechanism to initiate a dialogue. DIalogue is initiated so that agent's internal incoherence is reduced. At each step of this argumentation process coherence is reevaluated. However there are a number of ways our approach differ from theirs. First we treat the coherence framework from a more fundamental perspec-

tive by making coherence graphs corresponding to BDI modalities elementary. Thus we now have a clear way of studying the interactions among and between the cognitions whereas they have a very problem specific formulation of coherence. This also implies we can derive the associations between elements (constraints) from the properties of the underlying logic whereas they have no way of deriving these constraints. And at a broader level, we try introduce agent autonomy which is lacking in the current BDI models. Finally there is no work which gives a coherence framework to reason about agents and institutions, individually and together.

And finally the collection of works by Thagard who proposed the coherence theory as constraint satisfaction [11]. He has applied his theory to explain many of the natural phenomena. But so far has not given a formal specification of coherence nor integration into other theories.

6 Discussion and Future work

In this paper, we have formally defined the basic coherence tools for building institutional agents. We aim to further develop this theory in the following directions.

An important question we have left unanswered in the paper is given the beliefs or norms how their corresponding coherence graphs can be created. Evaluating the association between two atomic beliefs looks more like a human task, yet we can use similarity measures extracted from other repositories like ontologies, Wordnet or search results. Whereas evaluating associations between complex beliefs, we can use the underlying logic. Composing coherence graphs is another important aspect that we have dealt only superficially. The composition is important as it is the coherence measures of the graph compositions that normally identifies conflicts. We plan to explore these ideas in more detail in our future work.

In this paper we also limit our framework to logical systems whereas coherence can be applied to arbitrary graphs. In the future work we plan to make the coherence graphs more general so that non-logical agents can use coherence measures.

In the present work, we have provided the basic reasoning tools for a norm aware agent. We have shown when and how an autonomous agent could violate a norm. From the institutional perspective, a series of norm violations should trigger further actions, such as an analysis of why the norm is being violated. This could lead to a norm revision leading to an institutional redefinition. Our future work involves further exploration into questions related to norm violation from an institutional perspective.

We have simplified the representation of norms in the present work. In the future, we plan to have a more expressive representation of norms which includes the state of action when the norm is applicable, objectives behind the norm and the values promoted by the norm, borrowing the ideas developed in [1].

And finally, a coherence maximization may not only lead to a norm violation, but can also trigger a belief update, leading to the process of evolution of cognition. There are no widely accepted theories on how a cognitive agent can be evolved. The proposed theory helps to understand when a belief revision is profitable. In the future work, we propose to further explore cognitive revision in an institutional agent.

Acknowledgments. This work is supported under the OpenKnowledge[12] Specific Targeted Research Project (STREP), which is funded by the European Commission under contract number FP6-027253. Schorlemmer is supported by a *Ramon y Cajal* research fellowship from Spain's Ministry of Education and Science, which is partially funded by the European Social Fund. Special acknowledgments to all the reviewers of COIN@DURHAM07 for their detailed reviews and insightful comments.

References

1. K. Atkinson. *What Should We Do?: Computational Representation of Persuasive Argument in Practical Reasoning.* PhD thesis, University of Liverpool, 2005.
2. Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *AGENTS '01*, 2001.
3. A. Casali, L. Godo, and C. Sierra. Graded BDI models for agent architectures. In *lecture notes in computer science*, volume 3487, 2005.
4. Cristiano Castelfranchi, Frank Dignum, Catholijn M. Jonker, and Jan Treur. Deliberative normative agents: Principles and architecture. In *ATAL '99*, 2000.
5. Leon Festinger. *A theory of cognitive dissonance*. Stanford University Press, 1957.
6. Lou Goble and John-Jules Ch. Meyer. Deontic logic and artificial normative systems. In *DEON 2006*, 2006.
7. Justin Yifu Lin. An economic theory of institutional change: induced and imposed change. *Cato Journal*, 9(1), 1989.
8. John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Science Editions, J. Wiley, 1964.
9. Philippe Pasquier and Brahim Chaib-draa. The cognitive coherence approach for agent communication pragmatics. In *AAMAS '03*, 2003.
10. John R. Searle. *The Construction of Social Reality*. Free Press, 1997.
11. Paul Thagard. *Coherence in Thought and Action*. MIT Press, 2002.
12. Francesco Vigan, Nicoletta Fornara, and Marco Colombetti. An operational approach to norms in artificial institutions. In *AAMAS '05*, 2005.
13. G. H. von Wright. *An Essay in Deontic Logic and the General Theory of Action : with a Bibliography of Deontic and Imperative Logic*. North-Holland Pub. Co, 1968.
14. Fabiola López y López, Michael Luck, and Mark d'Inverno. Constraining autonomy through norms. In *AAMAS '02*, 2002.

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