Towards a Normative BDI Architecture for Norm Compliance

N. Criado[1], E. Argente[1], P. Noriega[2], and V. Botti[1]

[1] Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n. 46022 Valencia (Spain)
Email: {ncriado,eargente,vbotti}@dsic.upv.es
[2] Institut d'Investigació en Intel.ligència Artificial
Consejo Superior de Investigaciones Científicas
Campus de la UAB, Bellaterra, Catalonia (Spain)
Email: pablo@iiia.csic.es

Abstract. Multi-Agent Systems require coordination mechanisms in order to assemble the behaviour of autonomous and heterogeneous agents and achieve the desired performance of the whole system. Norms are deontic statements employed by these coordination mechanisms which define constraints to the potential excesses of agents' autonomous behaviour. However, norms are only effective if agents are capable of understanding and managing them pragmatically. In this paper, we propose an extension of the BDI proposal in order to allow agents to take pragmatic autonomous decisions considering the existence of norms. In particular, coherence and consistency theory will be employed as a criterion for determining norm compliance.

1 Introduction

The development of network technologies and Internet has made it possible to evolve from monolithic and centralized applications, in which problems are solved by a single component, to distributed applications, in which problems are solved by means of the interaction among autonomous agents. In these systems, the autonomy and heterogeneity of agents make mandatory the definition of *coordination* mechanisms for ensuring the whole performance of the system. With this aim, social notions, such as *organizations*, *institutions* and *norms*, have been introduced in the design and implementation of distributed systems.

Norms have been defined in distributed systems as regulations or patterns of behaviour established in order to constrain the potential excesses of autonomous agents. The definition of norms for controlling distributed systems requires the development of normative agents. Normative agents [8] must be endowed with capabilities for considering norms and deciding which norms to comply with and how to comply with them. The multi-context Graded BDI architecture [7] allows agents to reason in uncertain and dynamic environments.

The work presented in [9] is a first effort on the extension of the Graded BDI architecture [7] in order to allow agents to accept norms autonomously. This work focuses on the description of the architecture as a whole but it provides few details about how agents acquire new norms and face with the norm compliance dilemma. In addition, it lacks an elaborated definition of norm and norm dynamics. According to these criticisms, in this paper we propose to revise this architecture in order to allow agents to take norms into account in a more sophisticated way. In particular, here we focus on the application of both *Cognitive Coherence Theory* [26] and *Consistency Theory* [2] for reasoning about norm compliance. Coherence is a cognitive theory whose main purpose is the study of how pieces of information influence each other by imposing a positive or negative constraint over the rest of information. Consistency is a logic property which analyses the relationship among a formula and its negation. Our proposal consists on applying deliberative coherence theory for determining which norms are more coherent with respect to the agent's mental state. In addition, consistency criterion is considered when determining how to comply with norms. Therefore, this paper tries to overlap some of the main drawbacks of the original proposal by means of adding coherence and consistency constraints to the architecture.

This paper is structured as follows: next section describes the background of our proposal; Section 3 provides norm definitions; in Section 4 the normative BDI architecture is explained; the two components in charge of norm management are explained in Sections 5 and 6; Section 7 describes the norm internalization process; and, Section 8 remarks the contributions and future work.

2 Background

Along this section all approaches considered for this work are explained. In particular, the normative multi-context Graded BDI architecture (n-BDI for short) refined in this paper is explained first. Next subsections introduce the basis of consistency and coherence theories.

2.1 BDI architectures for normative agents

Usually, proposals on agent architectures which support normative reasoning do not consider norms as dynamic objects which may be acquired and recognised by agents. On the contrary, these proposals consider norms as static constraints that are hard-wired on agent architectures. Regarding recent proposals on individual norm reasoning, the BOID architecture [4] represents obligations as mental attributes and analyses the relationship and influence of such obligations on agent beliefs, desires and intentions. However, this proposal presents some drawbacks: i) it only considers obligation norms; ii) it considers norms as static entities that are off-line programmed in agents. In relation with this last feature, the EMIL proposal [1] has developed a framework for autonomous norm recognition. Thus, agents would be able to acquire new norms by observing the behaviour of other agents which are situated in their environments. The main disadvantage of EMIL

is that agents obey all recognised norms blindly without considering their own motivations. The multi-context graded BDI agent architecture [7] does not provide an explicit representation of norms. However, it is capable of representing and reasoning with graded mental attitudes, which makes it suitable as a basis for a norm aware agent architecture.

In order to overlap these drawbacks, in [9, 10], the multi-context graded BDI agent architecture [7] has been extended with recognition and normative reasoning capabilities. According to the n-BDI proposal, an agent is defined by a set of interconnected contexts, where each context has its own logic (i.e. its own language, axioms and inference rules). In addition, bridge rules are inference rules whose premises and conclusions belong to different contexts. In particular, an n-BDI agent [9, 10] is formed by:

- *Mental* contexts [6] to characterize beliefs (BC), intentions (IC) and desires (DC). These contexts contain logic propositions such as $(\Psi\gamma, \delta)$; where Ψ is a modal operator in $\{B, D^+, D^-, I\}$ which express beliefs, positive and negative desires and intentions, respectively; $\gamma \in \mathcal{L}_{\mathcal{DL}}$ is a dynamic logic [19] proposition; and $\delta \in [0, 1]$ represents the certainty degree associated to this mental proposition. For example $(B\gamma, \delta)$ represents a belief about proposition γ of an agent and δ represents the certainty degree associated to this belief.
- *Functional* contexts [6] for planning (PC) and communication (CC).
- *Normative* contexts [9] for allowing agents to recognise new norms (RC) and to consider norms in their decision making processes (NC).
- *Bridge Rules* for connecting mental, functional and normative contexts. A detailed description of these bridge rules can be found in [7]. For a more detailed description of normative bridge rules see [9].

Regarding the normative extension of the BDI architecture, the norm decision process consists of the following steps:

1. It starts when the RC derives a new norm through analysing its environment.
2. These norms are translated into a set of inference rules which are included into the NC. The NC is responsible for deriving new beliefs and desires according to the current agent mental state and the inference rules which have been obtained from norms.
3. After performing the inference process for creating new beliefs and desires derived from norm application, the normative context must update the mental contexts.

The original proposal [9, 10] is a preliminary work towards the definition of autonomous norm aware agents capable of making a decision about norm compliance. In this sense, this approach presents several problems and deficiencies. Firstly, the notion of norm is vague and imprecise, in this sense there is not a clear definition of what an abstract norm and a norm instance mean. Regarding the norm recognition process, no details about how the set of abstract norms is updated and maintained are provided. Thus, the RC is seen as a black box that

gives no analysis of how it deals with different types of norms (e.g. social norms, explicit norms created by the institution). Finally, it lacks a more concrete description of how a BDI agent may decide about obeying or not a norm. In this sense, the derivation of positive and negative desires from obligations and prohibitions is too simple. In particular, norms that guide agent behaviours might be in conflict, since they are aimed at defining the ideal behaviour of different roles which may be played by one agent. Besides that, norm compliance decisions should be consistent with the mental state of agents. Therefore, how agents make consistent decisions about norm compliance is the main contribution of the current paper with respect to the original proposal [9].

As a solution to this problem we will employ works on formalisms for ensuring *consistency* [2] and *coherence* [26]. In particular, this paper describes how these works are applied for reasoning about norm compliance. Next subsections briefly describe both the proposal of Casali et al. [6] on consistency among graded bipolar desires and the work of Joseph et al. [17] on the formalization of deductive coherence for multi-context graded BDI agents.

2.2 Consistency for Graded BDI Agents

In [2] Benferhat et al. made a study of consistency among bipolar graded preferences. Taking this definition of consistency, Casali et al. in [6] proposed several schemas for ensuring consistency among mental graded propositions. In particular, the maintenance of consistency among desires is achieved by means of three different schemas (i.e. DC_1, DC_2 and DC_3) which impose some constraints between the positive and negative desires of a formula and its negation. Thus, DC_2 schema (which will be employed in this paper) imposes a restriction over positive and negative desires for the same goal ($(D^+\ \gamma, \delta^+_\gamma)$ and $(D^-\ \gamma, \delta^-_\gamma)$, respectively). In particular, it claims that an agent cannot desire to be in world more than it is tolerated (i.e. not rejected). Therefore, it determines that:

$$\delta^+_\gamma + \delta^-_\gamma \leq 1$$

where δ^+_γ and δ^-_γ are the desirability and undesirability degrees (i.e. the certainty of the positive and negative desire) of proposition γ, respectively.

2.3 Coherence for Graded BDI Agents

In [26] Thagard claims that coherence is a cognitive theory whose main purpose is the study of associations; i.e. how pieces of information influence each other by imposing a positive or negative constraint over the rest of information. According to Thagard's formalization, a coherence problem is modelled by a graph $g = \langle V, E, \zeta \rangle$; where V is a finite set of nodes representing pieces of information, E are the edges representing the positive or negative constraints among information; each constraint has a weight ($\zeta : E \rightarrow [-1, 1] \setminus \{0\}$) expressing the constraint strength. Maximizing the coherence [26] is the problem of partitioning nodes

into two sets (accepted A and rejected $V \setminus A$) which maximizes the strength of the partition, which is the sum of the weights of the satisfied constraints.

Taking a proof-theoretic approach, Joseph et al. [17] provide a formalization of deductive coherence for multi-context graded BDI agents. Thus, this work proposes a formalization together with mechanisms for calculating the coherence of a set of graded mental attitudes. The main idea beyond this formalism is to consider the inference relationships among propositions belonging to the same context for calculating the weight of coherence and incoherence relationships. Similarly, bridge rules are employed for setting the coherence degree among propositions belonging to different contexts. Details concerning building the coherence graph can be found in [17].

Regarding the relation of coherence with normative decision processes, in [18] Joseph et al. employed coherence as a criterion for rejecting or accepting norms. However, this work is based on a very simple notion of norm as an unconditional obligation. Moreover, this proposal only considers coherence as the one rational criterion for norm acceptance. In addition, the problem of norm conflict has not been faced. Finally the process by which agents' desires are updated according to norms have also been defined in a simple way without considering the effect of these normative desires on the previous existing desires.

3 Norm Notion

Norms have been studied from different fields such as philosophy, psychology, law, etc. MAS research has given different meanings to the norm concept, been employed as a synonym of obligation and authorization [14], social law [20], social commitment [24] and other kinds of rules imposed by societies or authorities.

In this work, we take as a basis the formalization of norms made in [21]. In this proposal a distinction among *abstract norms* and *norm instances* is made. An *abstract norm* is a conditional rule that defines under which conditions obligations, permissions and prohibitions should be created. In particular, the activation condition of an abstract norm defines when an obligation, permission or prohibition must be instantiated. The *norm instances* that are created out of the *abstract norms* are a set of active unconditional expressions that bind a particular agent to an obligation, permission or prohibition. Moreover, a norm instance is accompanied by an expiration condition which defines the validity period or deadline of the norm instance.

Following this proposal our definition of both abstract norms and norm instances is provided.

Definition 1 (Abstract Norm). *An abstract norm is defined as a tuple* $n_a = \langle D, A, E, C, S, R \rangle$ *where:*

- $D \in \{\mathcal{F}, \mathcal{P}, \mathcal{O}\}$ *is the deontic type of the norm. In this work obligations* ($\mathcal{O}$) *and prohibitions* ($\mathcal{F}$) *impose constraints on agent behaviours; whereas permissions* ($\mathcal{P}$) *are operators that define exceptions to the activation of obligations or prohibitions;*

- *A is the norm activation condition. It defines under which circunstancies the abstract norm is active and must be instantiated.*
- *E is the norm expiration condition, which determines when the norm no longer affects agents.*
- *C is a logic formula that represents the state of affairs or actions that must be carried out in case of obligations, or that must be avoided in case of prohibition norms.*
- *S, R are expressions which describe the actions (sanctions S and rewards R) that will be carried out in case of norm violation or fulfilment, respectively.*

Since this work is focused on the norm compliance problem, only those norms addressed to the agent will be taken into account.

Definition 2 (Norm Instance). *Given belief theory Γ_{BC} an abstract norm $n_a = \langle D, A, E, C, S, R\rangle$ is instanciated into a norm instance $n_i = \langle D, C'\rangle$ where:*

- *$\Gamma_{BC} \vdash \sigma(A)$, where σ is a substitution of variables in A such that $A' = \sigma(A)$ and $\sigma(S)$, $\sigma(R)$ and $\sigma(E)$ are fully grounded.*
- *$C' = \sigma(C)$.*

Once the activation conditions of an abstract norm hold it becomes active and several norm instances, according to the possible groundings of the activation condition, must be created. For simplicity, we assume that once a norm is being instantiated then it is fully grounded. In our proposal, the instantiation of activation and expiration conditions are considered by the *Norm Instantiation* bridge rule (which will be explained in Section 6). Similarly, sanctions and rewards are also considered by this bridge rule in order to decide about convenience of norm compliance. Thus, for simplicity we omit the instantiation of the norm expiration and activation conditions ($\sigma(A)$ and $\sigma(E)$) and the sanction and reward ($\sigma(S)$ and $\sigma(R)$) in the representation of a norm instance.

4 Normative BDI Architecture (n-BDI)

As previously mentioned, the main contribution of this paper is to refine the n-BDI architecture, which was originally proposed in [9, 10], with a more elaborated notion of norm and norm reasoning. In order to design this second version of the n-BDI, the work of Sripada et al. [25] has been considered as a reference. It analyses the psychological architecture subserving norms. In particular, this architecture is formed by two closely linked innate mechanisms: one responsible for norm acquisition, which is responsible for identifying norm implicating behaviour and inferring the content of that norm; and the other in charge of norm implementation, which maintains a database of norms, detects norm violations and generates motivations to comply with norms and to punish rule violators.

The evolution of n-BDI is focused on reasoning about norm compliance and acceptance, so issues related to the detection and reaction to norm violation are beyond the scope of this paper. In this sense, norms affect n-BDI agents in

two ways: i) when a norm is recognised and accepted then it is considered to define new plans; and ii) when accepted norms are active then their instances are used for selecting the most suitable plan which complies with norms. This paper tackles with this last effect of norms. In particular, this paper describes how *Deductive Coherence* (described in Section 2.3) and *Consistency Theory* (described in Section 2.2) are applied for reasoning about norm compliance.

The n-BDI refines the normative contexts (described in Section 2.1) according to the norm notions introduced in Section 3. Therefore, Figure 1 shows a scheme of the n-BDI proposed in this paper. In particular, the RC has been redefined as the *Norm Acquisition Context* (NAC), whereas the NC has been redefined into the *Norm Compliance Context* (NCC).

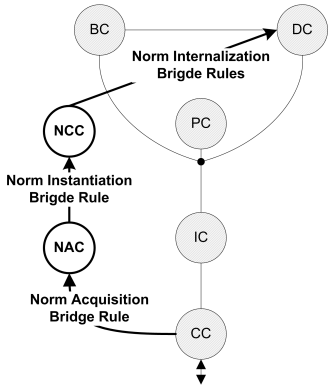

Fig. 1. Normative Extension of the Multi-Context BDI Architecture. Grey contexts and dashed lines (bridge rules) correspond to the basic definition of a BDI agent. The normative extensions are the white contexts and bold lines.

In this new version of the agent architecture not only the normative contexts have been improved by considering more elaborated normative definitions, but also the norm reasoning process has been extended with consistency and coherence notions. Thus, the norm reasoning process can be described as follows:

1. It starts when the NAC receives information cues for inferring new abstract norms through the *Norm Acquisition* bridge rule. The NAC carries out an inference process for maintaining the set of abstract norms in force in a specific moment.
2. Once the norm activation conditions hold, abstract norms are instantiated and included into the NCC by means of the *Norm Instantiation* bridge rule. Then, the NCC carries out an internal process for determining compliance with which of the norm instances. In this sense, not all active norms should be considered when updating the mental state. In this sense, our proposal consists in employing coherence theory as a criterion for determining *which* norms comply with. Therefore, the coherence maximization process is cal-

culated in order to determine which norm instances are consistent and must be taken into account when updating the desire theory.
3. Then, *Norm Internalization* bridge rules derive new desires according to the current agent mental state and the set of complied norms, also taking into account consistency considerations. These new desires may help the agent to select the most suitable plan to be intended and, as a consequence, normative actions might be carried out by the agent.

Thus, the norm reasoning process is formed by four different phases: acquisition of norms in force, decision about norm compliance and internalization of norms. Next sections describe each one of these phases in detail.

5 Norm Acquisition (NAC)

The NAC context allows agents to maintain a norm base that contains those norms which are *in force* in a specific moment (i.e. all norms which are currently applicable). Thus it is responsible for acquiring new norms and deleting obsolete norms; and updating the set of in force norms accordingly. This process can be defined as objective since no motivation or goal is considered in the acquisition process. Thus, agents only take into account their knowledge of the world in order to determine the set of norms which is more likely to be in force.

NAC Language. The NAC is formed by expressions such as (n, ρ) where n is an abstract norm according to Definition 1; and $\rho \in [0, 1]$ is a real value which assigns a degree to this abstract norm. This parameter ρ can have different interpretations. It can be defined as the reputation of the informer agent in case of leadership-based norm spreading. If norms are inferred by imitation, ρ might represent the acceptance degree of the norm. In case of utility maximizing approaches, as learning algorithms, it can be defined as the expected utility of the norm.

Abstract Norm Recognition. Regarding how new and obsolete norms are recognised, the NAC consists of a computational model of autonomous norm recognition which receives the agent perceptions, both observed and communicated facts, and identifies the set of norms which control the agent environment. Perceptions which are relevant to the norm recognition may be classified into:

- *Explicit normative perceptions.* They correspond to those messages exchanged by agents in which norms are explicitly communicated. Following this approach, several works have focused on analysing the role of leaders in the norm spreading. In particular, these leaders provide normative advices to follower agents when deciding about a norm [27, 22].
- *Implicit normative perceptions.* This type of perceptions includes the observation of actions performed by agents as a way of detecting norms. Since norms are usually supported by enforcing mechanisms such as sanctions and

rewards, the detection of them has been considered as an alternative for acquiring new norms [15]. Other works have proposed imitation mechanisms as a criterion for acquiring new norms. These models are characterized by agents mimicking the behaviour of what the majority of the agents do in a given agent society [28, 5]. Moreover, in [23] researchers have experimented with learning algorithms to identify a norm that maximizes an agent's utility.
- *Mixed normative perceptions.* There are proposals which consider both explicit and implicit normative perceptions as cues for inferring norms [1].

Abstract Norm Dynamics. The set of norms which are in force may change both explicitly, by means of the addition, deletion or modification of the existing norms; and implicitly by introducing new norms which are not specifically meant to modify previous norms, but which change in fact the system because they are incompatible with such existing norms and prevail over them [16]. However, this is a complex issue which is out of the scope of this paper. Works presented at the *Formal Models of Norm Change*[3] are good examples of proposals which provide a formal analysis of all kinds of dynamic aspects involved in systems of norms.

This paper does not focus on the norm acquisition problem and the dynamics of abstract norms. In the following, the NAC will be considered as a *black box* that receives cues for detecting norms as input and generates abstract norms as output.

6 Norm Compliance (NCC)

The NCC is the component responsible for reasoning about the set of norms which hold in a specific moment. In this sense the NAC recognises all norms that are in force, whereas the NCC only contains those norms which are active according to the current situation. The NCC should determine which and how norms will be obeyed and support agents when facing with norm violations. In this sense, the NCC detects norm violations and fulfilments and generates punishing and rewarding reactions. This last issue is over the scope of this paper and will be analysed in future works.

The functionalities carried out by the NCC which are covered by this work are related to three main issues: the NCC is in charge of maintaining the set of instantiated norms which are active; then it considers convenience of norm compliance and determines which norms comply with; and, finally, it derives new desires for fulfilling these norms.

NCC Language. The NCC is formed by expressions such as: (n, ρ) where n is a norm instance according to Definition 2. $\rho \in [0, 1]$ is a real value which assigns a degree to this norm instance. This parameter can be interpreted as the salience of the norm instance. Its value can be determined according to different

[3] http://www.cs.uu.nl/events/normchange2/

criteria such as utility of norm compliance, emotional considerations, intrinsic motivations, etc. In this paper, it is defined with regard to the certainty of norm activation as well as the convenience of norm compliance.

Instantiated norms are inferred by applying instantiation bridge rules to norms when their activation is detected. Next, these normative bridge rules are described in detail.

Norm Instantiation Bridge Rule.

$$\frac{NAC : (\langle D, A, E, C, S, R\rangle, \rho), BC : (B\,A, \beta_A), BC : (B\neg E, \beta_E)}{NCC : (\langle D, C\rangle, f_{instantiation}(\theta_{activation}, \theta_{compliance}))} \quad (1)$$

If an agent considers that an abstract norm $n_a = \langle D, A, E, C, S, R\rangle$ is currently active $((B\,A, \beta_A) \wedge (B\neg E, \beta_E))$ then a new norm instance $n_i = \langle D, C\rangle$ is generated. The degree assigned to the norm instance is defined by the $f_{instantiation}$ function which combines the values obtained by the $\theta_{activation}$ and $\theta_{compliance}$ functions.

On the one hand, $\theta_{activation}$ combines the evidence about norm activation (i.e. the certainty degrees β_A, β_E and ρ). It can be given a sophisticated definition depending on the concrete application. In this work, it has been defined as the weighed average among these three values, as follows:

$$\theta_{activation} = \frac{w_A \times \beta_A + w_E \times \beta_E + w_\rho \times \rho}{w_A + w_E + w_\rho}$$

If all values are equally weighed, then we obtain that $\theta_{activation} = \frac{\beta_A + \beta_E + \rho}{3}$

On the other hand, $\theta_{compliance}$ considers both intrinsic and instrumental motivations for norm compliance. In [11] different strategies for norm compliance from an instrumental perspective over this architecture are described. In particular, they consider the influence of norm compliance and violation on agent's goals for determining whether the agent accepts the norm. For example, an *egoist* agent will accept only those norms which benefit its goals (i.e. whose condition is positively desired). In this case:

$$\theta_{compliance} = \begin{cases} 1 & \text{if } \delta_C^+ > 0, \text{ where } (D^+C, \delta_C^+) \in \Gamma_{DC}; \\ 0 & \text{otherwise} \end{cases}$$

Finally, values obtained by the $\theta_{activation}$ and the $\theta_{compliance}$ functions are combined by the $f_{instantiation}$:

$$f_{instantiation}(\theta_{activation}, \theta_{compliance}) = \frac{w_{activation} \times \theta_{activation} + w_{compliance} \times \theta_{compliance}}{w_{activation} + w_{compliance}}$$

Again, if these two parameters are equally weighed, then we obtain that

$$f_{instantiation}(\theta_{activation}, \theta_{compliance}) = \frac{\theta_{activation} + \theta_{compliance}}{2}$$

This approach relies upon various values such as $w_{compliance}$, w_A and $w_{activation}$. The definition of these values is beyond the scope of this paper. In previous works [9, 11, 10], it has been considered that they are defined off-line by the agent designer. However, this solution is static and it does not allow agents to adapt these values according to a changing environment. Thus, this issue will be considered in future works.

6.1 Coherence For Norm Instances

Once *Norm Instantiation* bridge rule has been applied, it is possible that there is an incoherence between mental propositions. Because of this, a maximizing coherency process is needed in order to determine which propositions are consistent and must be taken into account; and which propositions belonging to the rejection set will be ignored when deriving normative desires.

Since our proposal of agent architecture employs graded logics for representing mental propositions, this work takes as a basis the work described in Section 2.3. As argued before, this work proposes a formalization together with mechanisms for calculating the deductive coherence of a set of graded mental attitudes. Our proposal adapts this work by applying the coherence maximization algorithm to the norm compliance problem. Figure 2 illustrates an overview of the employment of coherence as a criterion for resolving the norm compliance dilemma. As shown by this figure, the normative coherence process considers propositions belonging to the BC, the NCC and NAC. Basically this process takes into account: i) the beliefs that sustain the activation of norms and their relationships among them and other beliefs of the BC; ii) the norm instances and conflict relationships among them; and iii) the abstract norms which have triggered the norm activation. Relationships among propositions belonging to each context are defined by means of inference rules and axioms, whereas coherence connections among propositions of different contexts are defined by means of norm instantiation bridge rules.

By considering coherence we will address three different problems: i) determining norm deactivation; ii) determining active norms in incoherent states and iii) normative conflict resolution. In order to formalize normative incoherence the original proposal of [17] must be extended with extra constraints. Moreover, since we apply the coherence calculation algorithms for improving the normative reasoning, then only those propositions which are relevant to the norm compliance process are taken into account. Next, both the definition of normative coherence constraints for facing with each one of these three questions as well as the determination of relevant propositions is detailed.

Detecting Norm Activation in Incomplete and Inconsistent States. As illustrated in Figure 2, norm instantiation bridge rule (see equation 6) allows norm instances (from NCC) to be connected to beliefs from BC related to their activation and expiration conditions. Norm instantiation bridge rule has as preconditions the belief about the occurrence of the activation condition A and the

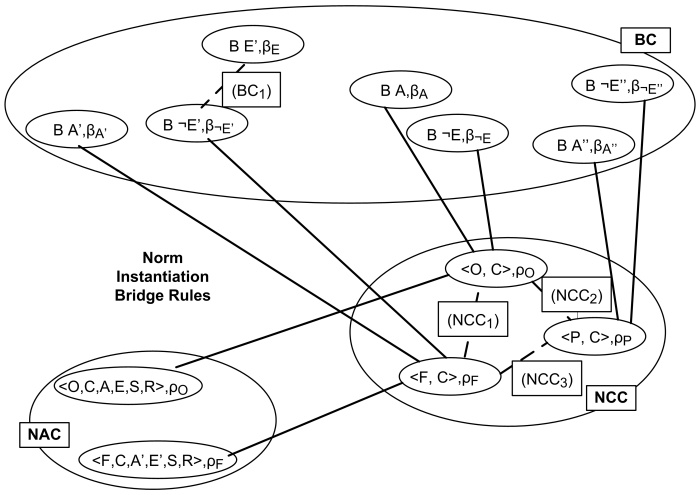

Fig. 2. Usage of coherence as a criterion for resolving the norm compliance dilemma.

negation of the expiration condition E. Usually agents do not have an explicit knowledge about the negation of E. However, it is possible to infer a certainty degree about $\neg E$ from the certainty degree of E. Following this idea, the first step for computing coherence is to calculate the closure under negation of beliefs as follows:

Definition 3 (Closure of Beliefs under Negation). *Let Γ be a finite belief theory presentation using graded formulas. We define the closure of Γ under negation as:*

$$\Gamma^{\neg} = \Gamma \cup \{(B\neg\varphi, (1-\delta)) : (B\neg\varphi, \beta) \notin \Gamma \text{ and } (B\varphi, \delta) \in \Gamma\}$$

Therefore, the closure of a set of beliefs under negation consists on extending this theory by inferring new information from what is actually believed. In particular, if an agent believes that proposition E is true with a certainty degree δ but it does not have any belief concerning its negation, it is logic to assume that the certainty degree assigned to $\neg E$ should be lower than $(1-\delta)$. We need to calculate the closure of beliefs under negation in order to detect norm deactivation. In this sense, when the certainty about the expiration condition E increases it can be inferred that the certainty of $\neg E$ decreases even if the agent does not have explicit evidence of it.

In addition we want to define an incoherence relationship among a belief related to a general proposition and its negation. This relationship is defined by means of the addition of an inference rule in the belief context:

$$(BC_1)\ (B\gamma, \beta_\gamma), (B\neg\gamma, \beta_{\neg\gamma}) \vdash (\bar{0}, 1-(\beta_\gamma + \beta_{\neg\gamma}))$$

Basically this scheme means that to belief γ and $\neg\gamma$ simultaneously is a contradiction $(\bar{0})$. The certainty degree of this contradiction is defined in [18] as $1-(\beta_\gamma + \beta_{\neg\gamma})$.

One of the main problems of the multi-context BDI architecture is the fact that it does not allow the definition of bridge rules for deleting propositions. In this sense, there is a bridge rule for inferring a new instance of a norm when its activation condition holds. However, it is not possible to create a bridge rule which deletes this instance when the expiration condition holds. In response to this problem, coherence will be used as a criterion for detecting norm deactivation. Moreover, an agent may have beliefs related to the occurrence of both the norm activation and expiration conditions. Thus it should consider all those evidences that sustain the occurrence of the expiration and activation conditions in order to determine the set of norms which are active. In particular, coherence will be used as a criterion for detecting norm activation/deactivation according to the certainty of both the expiration and the activation conditions.

Resolving Normative Conflicts. As previously argued, the above process of normative coherence is useful not only to determine which norms are active but even to resolve a norm conflict. Usually, a norm conflict has been defined in other works as a situation in which something is considered as forbidden and obliged or forbidden and permitted. In our proposal, we define permissions as a normative operator which allows defining an exception to the application of a more general obligation or prohibition norm. Thus, we also consider that norms which define something as forbidden and permitted are also in conflict. However, there is no constraint that represents this type of incoherence. In order to represent incoherence inferred from norm conflicts we add the next inference rules to the NCC:

$$
\begin{array}{ll}
(NCC_1) & (\langle O, C\rangle, \rho_O), (\langle F, C\rangle, \rho_F) \vdash (\bar{0}, -min(\rho_O, \rho_F)) \\
(NCC_2) & (\langle O, C\rangle, \rho_O), (\langle P, \neg C\rangle, \rho_P) \vdash (\bar{0}, 1 - (\rho_O + \rho_P)) \\
(NCC_3) & (\langle F, C\rangle, \rho_F), (\langle P, C\rangle, \rho_P) \vdash (\bar{0}, 1 - (\rho_F + \rho_P))
\end{array}
$$

In case of a conflict between a permission and an obligation or a prohibition, the degree of the falsity constant ($\bar{0}$) is assigned value $1-(\rho_O+\rho_P)$ or $1-(\rho_F+\rho_P)$, respectively, in a similar way as in BC_1. In case of a conflict among a prohibition and an obligation we define a stronger incoherence by defining the degree of the falsity constant as $-min(\rho_O, \rho_F)$.

Selecting Relevant Propositions. Once the coherence graph has been defined and a maximising partition $(A, V \setminus A)$ over this graph has been found following [17], the set of propositions belonging to the NCC (i.e.Γ_{NCC}) is revised in order to consider only the accepted norms:

$$\Gamma'_{NCC} = \Gamma_{NCC} \cap A$$

where A is the accepted set of norm instances according to the maximizing coherence process [17], i.e. "the most coherent norm instances".

7 Norm Internalization

Regarding works on norm *internalization* in the MAS community, maybe the most relevant proposal are the works of Conte et al. [8]. According to them, a characteristic feature of norm internalization is that norms become part of the agent's identity. The concept of identity implies that norms become part of the cognitions of the individual agent. In particular, Conte et al. define norm internalization as a multi-step process, leading from externally enforced norms to norm-corresponding goals, intentions and actions with no more external enforcement. Thus they account for different types and levels of internalization.

In this paper a simplistic approximation to the norm internalization process has been considered. However, it will be object of future work extensions. In particular, we have only considered the internalization of norms as goals. In this sense, the process of norm internalization has been described by the self-determination theory [13] as a dynamic relation between norms and desires. This shift would represent the assumption that internalised norms become part of the agent's sense of identity. Thus, after performing the coherence process for creating new norm instances, the NCC must update the DC (Figure 1 Norm Internalization Bridge Rules) with the new normative desires. The addition of these propositions into this mental context may cause an inconsistency with the current mental state. As explained in Section 2.2, in [6] three schemas for ensuring consistency among mental graded propositions have been proposed. According to schema DC_2, which imposes a restriction over positive and negative desires for a same goal, we have implemented the following inference rule:

$$(DC_2)\ (D^+\gamma, \delta^+_\gamma), (D^-\gamma, \delta^-_\gamma) \vdash (\bar{0}, 1 - (\delta^+_\gamma + \delta^-_\gamma))$$

Our proposal needs bipolar representation of desires since it is useful when selecting plans to be intended for achieving the desires. In this sense, both negative and positive effects of actions will be taken into account when selecting a plan to be intended. For example, the fact that a plan involves a forbidden action may be considered as a negative effect. Therefore, obligation norms are internalized as positive desires whereas prohibition norms are translated into negative ones. Because of this, DC_2 has been considered as a basis for the definition of bridge rules responsible for updating the DC in a consistent way. Next these norm internalization bridge rules are described.

Norm Internalization Bridge Rules.

- *Obligation Norm.* According to DC_2 schema, bridge rule for updating the DC with the positive desires derived from obligation norms is defined as follows:

$$\frac{NCC : (\langle O, C\rangle, \rho), DC : (D^-\ C, \delta^-), DC : (D^+\ C, \delta^+)}{DC : (D^+\ C, max(\rho, \delta^+)), DC : (D^-\ C, min(\delta^-, 1 - max(\rho, \delta^+)))} \quad (2)$$

If an agent considers that the obligation is currently active then a new positive desire will be inferred corresponding to the new norm condition. Thus, the desire degree assigned to the new proposition C is defined as the maximum between the new desirability and the previous value ($max(\rho, \delta^+)$). Moreover, the undesirability assigned to C is updated as the minimum between the previous value of undesirability assigned to γ (δ^-) and its maximum coherent undesirability, which is defined as $1 - max(\rho, \delta^+)$.

- *Prohibition Norm.* Bridge rule for updating the DC with negative desires is defined as follows:

$$\frac{NCC : (\langle F, C\rangle, \rho), DC : (D^-\ C, \delta^-), DC : (D^+\ C, \delta^+)}{DC : (D^-\ C, max(\rho, \delta^-)), DC : (D^+\ C, min(\delta^+, 1 - max(\rho, \delta^-)))} \quad (3)$$

  Similarly to obligation norms, a prohibition related to a condition C is transformed into a negative desire related to the norm condition ($D^-\ C, max(\rho, \delta^-)$).

- *Permission Norm.* Finally, permission norms do not infer a positive or negative desire about the norm condition. Permission norms define exceptions to the application of a more general obligation or prohibition norm. As a consequence, they only are defined for creating an incoherence with these more general norms.

8 Conclusion

In this work a previous proposal [9, 10] of a normative BDI architecture has been revised. The first contribution of the current paper is the usage of coherence theory in order to determine what means to follow or violate a norm according to the agent's cognitions and making a decision about norm compliance. The second contribution of this paper is the employment of consistency notions for updating agent cognitions in response to these normative decisions.

The impact of normative decisions on agent cognitions will be object of future work. In this paper, the norm internalization problem has been faced in a simplistic way by considering only the impact of obeyed norms on agent's desires. Therefore, in future works the role of both *deliberative coherence* [26] and *emotions* on the norm compliance will be analysed. In particular, *deliberative coherence*, which deals with goal adoption in the context of decision making, will be considered when building plans for obeying or violating norms. In addition, we will work on extending our agent architecture with an explicit representation of emotions which will allow agents to consider phenomena such as shame, honour, gratitude, etc. in their decision making processes.

Due to lack of space, no evaluation or case study has been included here that might provide a more understanding perspective of our proposal. However, works describing the original proposal [10, 11] (neither consistency nor coherence are considered here) provide examples belonging to the m-Water case study. The m-Water [3] is a water right market which is implemented as a regulated open

multi-agent system. It is a challenging problem, specially in countries like Spain, since scarcity of water is a matter of public interest. The m-Water framework [12] is a somewhat idealized version of current water-use regulations that articulate the interactions of those individual and collective entities that are involved in the use of water in a closed basin. This is a regulated environment which includes the expression and use of regulations of different sorts: from actual laws and regulations issued by governments, to policies and local regulations issued by basin managers, and to social norms that prevail in a given community of users. For these reasons, we consider the m-Water problem as a suitable case study for evaluating performance of the n-BDI agent architecture, since agents' behaviour is affected by different sorts of norms which are controlled by different mechanisms such as regimentation, enforcement and grievance and arbitration processes.

Finally, we are working on the implementation of a prototype of the n-BDI architecture. Our aim is to evaluate empirically our proposal through the design and implementation of scenarios belonging to the m-Water case study. In future works, we will make some experiments concerning the flexibility and performance of the n-BDI agent model with respect to simple BDI agents, using the m-Water case study. However, preliminary results of the experimental evaluation of the n-BDI original proposal can be found in [10].

9 Acknowledgments

This work was partially supported by the Spanish government under grants CONSOLIDER-INGENIO 2010 CSD2007-00022, TIN2009-13839-C03-01 and TIN2008-04446 and by the FPU grant AP-2007-01256 awarded to N. Criado.

References

1. G. Andrighetto, M. Campenní, F. Cecconi, and R. Conte. How agents find out norms: A simulation based model of norm innovation. In *NORMAS*, pages 16–30, 2008.
2. S. Benferhat, D. Dubois, S. Kaci, and H. Prade. Bipolar representation and fusion of preferences on the possibilistic logic framework. In *KR*, pages 421–434. Morgan Kaufmann Publishers; 1998, 2002.
3. V. Botti, A. Garrido, A. Giret, and P. Noriega. Managing water demand as a regulated open mas. In *MALLOW Workshop on COIN*, page In Press., 2009.
4. J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre. The boid architecture – conflicts between beliefs, obligations, intentions and desires. In *AAMAS*, pages 9–16. ACM Press, 2001.
5. M. Campenní, G. Andrighetto, F. Cecconi, and R. Conte. Normal= Normative? The role of intelligent agents in norm innovation. *Mind & Society*, 8(2):153–172, 2009.
6. A. Casali, L. Godo, and C. Sierra. A logical framework to represent and reason about graded preferences and intentions. In *KR*, pages 27–37. AAAI Press, 2008.

7. A. Casali, L. Godo, and C. Sierra. *On Intentional and Social Agents with Graded Attitudes.* PhD thesis, Universitat de Girona, 2008.
8. R. Conte, G. Andrighetto, and M. Campenni. On norm internalization. a position paper. In *EUMAS*, 2009.
9. N. Criado, E. Argente, and V. Botti. A BDI Architecture for Normative Decision Making (Extended Abstract). In *9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, pages 1383–1384, 2010.
10. N. Criado, E. Argente, and V. Botti. Normative deliberation in graded bdi agents. In *MATES*, page In Press, 2010.
11. N. Criado, E. Argente, and V. Botti. Rational strategies for autonomous norm adoption. In *9th International Workshop on Coordination, Organization, Institutions and Norms in Multi-Agent Systems (COIN@AAMAS2010)*, pages 9–16, 2010.
12. N. Criado, E. Argente, A. Garrido, J. A. Gimeno, F. Igual, V. Botti, P. Noriega, and A. Giret. Norm enforceability in electronic institutions? In *11th International Workshop on Coordination, Organization, Institutions and Norms in Multi-Agent Systems (COIN@MALLOW2010)*, page In Press, 2010.
13. E. Deci and R. Ryan. The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4):227–268, 2000.
14. F. Dignum. Autonomous agents with norms. *Artif. Intell. Law*, 7(1):69–79, 1999.
15. F. Flentge, D. Polani, and T. Uthmann. Modelling the emergence of possession norms using memes. *Journal of Artificial Societies and Social Simulation*, 4(4), 2001.
16. G. Governatori and A. Rotolo. Changing legal systems: Abrogation and annulment part i: Revision of defeasible theories. In *DEON*, pages 3–18, 2008.
17. S. Joseph, C. Sierra, M. Schorlemmer, and P. Dellunde. Formalising deductive coherence: An application to norm evaluation. Technical report, IIIA-CSIC, 2009.
18. S. Joseph, C. Sierra, M. Schorlemmer, and P. Dellunde. Deductive coherence and norm adoption. *Logic Journal of the IGPL*, page In Press, 2010.
19. J. Meyer. Dynamic logic for reasoning about actions and agents. In *Logic-Based Artificial Intelligence*, pages 281–311. Kluwer Academic Publishers, 2000.
20. Y. Moses and M. Tennenholtz. Artificial social systems. *Computers and Artificial Intelligence*, 14(6), 1995.
21. N. Oren, S. Panagiotidi, J. Vázquez-Salceda, S. Modgil, M. Luck, and S. Miles. Towards a formalisation of electronic contracting environments. In *COIN IV*, pages 156–171, Berlin, Heidelberg, 2009. Springer-Verlag.
22. B. T. R. Savarimuthu, M. Purvis, and M. K. Purvis. Social norm emergence in virtual agent societies. In *AAMAS*, pages 1521–1524. IFAAMAS, 2008.
23. S. Sen and S. Airiau. Emergence of norms through social learning. In *IJCAI*, pages 1507–1512, 2007.
24. M. P. Singh. An ontology for commitments in multiagent systems. *Artif. Intell. Law*, 7(1):97–113, 1999.
25. C. Sripada and S. Stich. A framework for the psychology of norms. *The Innate Mind: Culture and Cognition*, pages 280–301, 2006.
26. P. Thagard. *Coherence in Thought and Action.* The MIT Press, Cambridge, Massachusetts, 2000.
27. H. Verhagen. *Norm Autonomous Agents.* PhD thesis, Stockholm University, 2000.
28. F. y Lopez. *Social Power and Norms.* PhD thesis, Citeseer, 2003.