

BDI+C — an architecture for normative, autonomous agents

Dorian Gaertner Pablo Noriega Carles Sierra

Institut d'Investigació en Intel·ligència Artificial, IIIA
Consejo Superior de Investigaciones Científicas, CSIC

Internal Report: RR-III A-2006-06

Abstract

In this paper, we describe a novel agent architecture for normative multi-agent systems which is based on multi-context systems. It models the three modalities of Rao and Georgeff's BDI agents as individual contexts and adds a fourth one for *commitments*. This new component is connected to all other mental attitudes via two sets of bridge rules, injecting formulae into it and modifying the BDI components after reasoning about commitments. As with other normative approaches the need for methods to deal with consistency is a key concern. We suggest three forms of dealing with the truth maintenance problem, all of which profit from the use of multi-context systems.

1 Introduction

Many researchers in the Artificial Intelligence community have identified autonomous agents as an important development towards the achievement of many of AI's promises. Among the many proposed agent architectures are Rao and Georgeff's BDI agents [26] that model mental attitudes of an agent, concretely *beliefs* (representing the state of the environment), *desires* (representing the state of affairs the agent wants to bring about) and *intentions* (representing the currently selected goals). Their architecture is well-known for its elegant logical formalisms, its foundation in philosophical theories and the many real-world applications that have been implemented using it (for example, dMARS [11]).

Multi-context systems allow to define complex systems with different formal components and the relationships between them. Giunchiglia and Serafini devised them a

decade ago in order to structure knowledge into distinct sets of facts or theories [16]. Parsons et al. [25] use these systems in order to model the three BDI modalities as individual components (*contexts* or *units*) with bridge rules to describe the dependencies between components. We propose to extend the BDI agent model with a fourth component that keeps track of the *commitments* an agent has adopted. We view a commitment as a triple consisting of the entity that commits, the entity that the commitment is directed at and the content of the commitment (similar to [5]). These entities can be individual agents, groups of agents or institutions. In this paper, we will follow the approach taken by Parsons et al. and model agents as multi-context systems. We describe how the commitment component is connected to the other three contexts via instances of two basic bridge rule schemata and suggest approaches to handle arising inconsistencies.

In the next section we are going to formally define the use of the terms *Commitment* and *Norm* that we are employing in this paper. Section 3 will introduce the multi-agent ballroom scenario that we are using to exemplify the normative aspects of our architecture in later sections. Subsequently, we briefly summarise multi-context systems and explain how we extended them. We show how our architecture lends itself to modelling *normative* MAS and propose a novel way to operationalise norms. Section 5 is concerned with truth maintenance and consistency issues and section 6 introduces norm adoption using the ballroom scenario. Finally, we contrast our architecture with existing proposals, present the open challenges, outline our future work and conclude.

2 Commitments and Norms

Norms, normative agents and normative multi-agent systems have received a lot of attention in recent years. López y López et al. [20] proposed a formal model of these concepts using the Z specification language. García-Camino et al. [14, 15] have analysed the concept of norms in a society of agents and how norms are implemented in an electronic institution. In [10], Dignum et al. extend the BDI architecture to handle norms. They are using PDL, a deontic logic, to formalise obligations from one agent to another. Norms, in their view, are obligations of a society or organisation. They explicitly state that a norm of a society is a conditional (p should be true when q is true). Finally, Cohen and Levesque in their seminal paper ‘Intention is choice with commitment’ [7] talk about internal commitments as a precursor to the social commitments that we concern ourselves with. Our view is more generic as described in this section.

We consider a *commitment* from one entity to be directed at another entity. With respect to these entities, one needs to distinguish between individual agents and groups of agents - or electronic institutions [1]. For example, an agent can be committed to an (electronic) institution to behave in a certain way. The institution on the other hand, may be committed to the agent to reward or punish him, depending on his behaviour. Note, that this is different from the case where one agent is responsible for norm en-

forcement. In such a case, one would have a commitment between the agent and the enforcer and another one between the enforcer and the agent. Commitments may also exist between agents or between different electronic institutions.

The content of a commitment can be a certain contract (e.g. an intention to deliver ten crates of apples once the agent believes to have been paid) between two agents or it can be a norm (e.g. you should not desire your neighbour's wife). In this paper, we will focus on the latter. The BNF description of our commitment language is hence:

$$\begin{aligned}
\textit{Commitment} &::= \textit{Commit}(S, S, \lceil C \rceil) \\
S &::= \textit{agent} \mid \textit{institution} \\
C &::= \textit{Contract} \mid \textit{Norm} \\
\textit{Contract} &::= \textit{WFF}^1 \\
\textit{Norm} &::= \varphi \rightarrow \psi \\
\varphi &::= \textit{ConjLiterals} \\
\textit{ConjLiterals} &::= \textit{MLiteral} \mid \textit{MLiteral} \wedge \textit{ConjLiterals} \\
\psi &::= \textit{MLiteral} \\
\textit{MLiteral} &::= \textit{MentalAtom} \mid \neg \textit{MentalAtom} \\
\textit{MentalAtom} &::= \textit{B}(\textit{term}) \mid \textit{D}(\textit{term}) \mid \textit{I}(\textit{term})
\end{aligned}$$

We consider a *norm* to be a conditional, first-order logic formula that relates mental attitudes of an agent. All variables are implicitly universally quantified. In this paper, we are using beliefs (B), desires (D) and intentions (I) to model an agent's mental state. For example,

$$\textit{B}(\textit{loves}(X, Y)) \rightarrow \neg \textit{I}(\textit{hurts}(X, Y))$$

is a norm which can be read as “for any two agents, if the agent X believes that it *loves* an agent Y then he should not intend to hurt her”. The argument to the mental literals can be any term and in a way, the implication arrow of the norm can always informally be translated with the English word *should*.

Although, in the above formula the modalities B and I should have a subscript x indicating that we are talking about beliefs and intentions of agent x we drop these subscripts for readability whenever it is clear from the context which agent is referred to. Furthermore, we never need distinct subscripts in the same norm formula, since it does not make sense to say a belief of one agent causes an intention for another agent.

A norm, for us, is a social phenomenon, in that it applies to all agents in a given society or institution. Each agent is then committed to the institution to obey the norm. We therefore stipulate that if φ is a norm in an institution Π , the following must hold:

¹A well-founded formula in an appropriate language, see e.g. [17].

$$\forall \alpha : agent \in \Pi : inst . Commit(\alpha, \Pi, \lceil \varphi \rceil)$$

where $\lceil \varphi \rceil$ is the codification of a norm as a term in Gödel's sense. Contrast this with the notion of a contract, which in most cases affects only two parties (or agents):

$$Contract(\alpha, \beta, \lceil \varphi \rceil) \rightarrow JointCommit(\{\alpha, \beta\}, \lceil \varphi \rceil)$$

where the notion of a joint commitment can be defined in arbitrarily complex ways. It could simply be a reciprocal commitment of the following kind:

$$JointCommit(\{\alpha, \beta\}, \lceil \varphi \rceil) \equiv Commit(\alpha, \beta, \lceil \varphi \rceil) \wedge Commit(\beta, \alpha, \lceil \varphi \rceil)$$

or it could involve more complicated notions such as (mutual) awareness of commitment and so on (for more work on joint commitments, see for example [8] and [23]). In what follows, we will mostly talk about agents who are committed to the institution they belong to. We therefore drop this information (i.e. the first two parameters) for brevity's sake. Unless otherwise stated, a commitment to φ of the form:

$$Commit(\lceil \varphi \rceil)$$

should be read as:

$$Commit(self, myInstitution, \lceil \varphi \rceil)$$

3 Ballroom Scenario

Throughout this paper, we are using the ballroom case-study developed by Gaertner et al. in [13]. They describe a normative multi-agent system simulation of a social ballroom with autonomous dancer agents that negotiate about and enter joint commitments to dance or drink together. The case study focuses on interaction protocols, norms and conventions that constrain the agents' behaviour. They also make these norms and conventions explicit, allowing for a more modular system design. Conventions of a ballroom can thus be changed dynamically or even evolve.

We are interested in particular in the second phase of the simulation, in which negotiations about future *dance commitments* take place following fixed, but varied protocols. Note that these dance commitments are distinct from the normative commitments (i.e.

commitments to norms) we described in the previous section. Unfortunately, the terminology is rather misleading and we will therefore be careful to be clear about which kind of commitment we are referring to.

Below we list a number of conventions, or normative commitments, that can exist in a certain social ballroom and restrict the behaviour of the participants:

1. dance partners should be of opposite sex
2. a female dancer should not approach a male dancer
3. one should not dance more than two consecutive dances with the same partner
4. one should dance at least once with one's mother-in-law if one desires to be a good husband

Expressed as norms using well-formed formulae built only from the modalities of belief, desire and intention, these read:

1. $\mathbb{B}(\text{sex}(X, S)) \wedge \mathbb{B}(\text{sex}(Y, S)) \rightarrow \neg \mathbb{I}(\text{danceWith}(X, Y))$
2. $\mathbb{B}(\text{sex}(X, \text{female})) \wedge \mathbb{B}(\text{sex}(Y, \text{male})) \rightarrow \neg \mathbb{I}(\text{initiateConversation}(X, Y))$
3. $\mathbb{B}(\text{lastDance}(X, Y)) \wedge \mathbb{B}(\text{penultimDance}(X, Y)) \rightarrow \neg \mathbb{I}(\text{danceWith}(X, Y))$
4. $\mathbb{D}(\text{goodHubby}(X)) \wedge \mathbb{B}(\text{momInLaw}(X, Y)) \wedge \mathbb{B}(\text{notYetDancedWith}(Y)) \rightarrow \mathbb{I}(\text{danceWith}(X, Y))$

One may argue, that these are standard bridge rules, however, the codification of these conditional formulae is used as an argument to or the content of a commitment. Remember, that the implication reads as *should*. The last example therefore *should* result in an intention to dance with the mother-in-law. There are many conceivable ways in which this intention fails to materialise, one example being a situation in which X is not committed to (desiring to) being a good husband.

An alternative representation of norms could take the notion of time into account. Consecutively performed dances, as mentioned in example 3 above, can be reasoned over using variants of the *Event Calculus* [18]. However, the other examples do not require explicit representation of time (e.g. dance partners should *always* be of opposite sex and a female dancer should *never* approach a male dancer). We therefore leave the inclusion of time into our norm representation for future work.

4 Multi-context Architecture: BDI+C

Multi-context systems have first been proposed a decade ago by Giunchiglia and Serafini in [16] and were subsequently used in a generic agent architecture by Noriega and Sierra in [22]. Individual theoretical components of an agent are modelled as separate *contexts* or units, each of which contains a set of statements in a language L_i together with the axioms A_i and inference rules Δ_i of a (modal) logic. A unit is hence a triple of the form:

$$unit_i = \langle L_i, A_i, \Delta_i \rangle$$

Not only can new statements be deduced in each context using the deduction machinery of the associated logic, but these contexts are also inter-related via *bridge rules*, that allow the deduction of a formula in one context based on the presence of certain formulae in other, linked contexts. An agent is then defined as a set of context indices I , a function that maps these indices to contexts, another function that maps these indices to theories (providing the initial set of formulae in each context) together with a set of bridge rules BR as follows:

$$Agent = \langle I, I \rightarrow \langle L_i, A_i, \Delta_i \rangle, I \rightarrow T_i, BR \rangle$$

We will now briefly summarise the main features of these systems, following the approach outlined by Parsons et al. in [25].

- *Units*: Structural entities representing the main components of the architecture— $unit_i$
- *Logics*: Declarative languages (L_i), each with a set of axioms (A_i) and a number of rules of inference (Δ_i). Each unit has a single logic associated with it.
- *Theories*: Sets of formulae written in the logic associated with a unit— T_i .
- *Bridge rules*: Rules of inference which relate formulae in different units— BR .

The BDI+C agent architecture we are proposing in this paper extends Rao and Georgeff's well-known BDI architecture with a fourth component which keeps track of the commitments of an agent. Below we describe a BDI+C agent as a multi-context system being inspired by the work of Parsons, Sierra and Jennings (see [25]). Each of the mental attitudes is modelled as an individual unit. Contexts for communication and planning (a functional context) are present in addition to the belief-, desire-, intention- and commitment-context but in this paper we will focus on the latter four. Each unit has its own logic associated with it.

For the belief context, we follow the standard literature (see for example, [24] and [26]) and choose the modal logic KD45 which is closed under implication, provides consistency, as well as positive and negative introspection. However, it does not have veridicality, which means that the agent’s beliefs may be false. However, in such a situation the agent itself is not aware that its beliefs are false. Like Rao and Georgeff [26] we also choose the modal logic KD to model the desire and intention components.

For the commitment context, the logic consists of the axiom schema K, closure under implication, together with the consistency axiom D. This does allow for conflicting commitments, but prohibits to be committed to something and not be committed to that something at the same time. That means, that we do not allow both $Commit(\lceil\varphi\rceil)$ and $\neg Commit(\lceil\varphi\rceil)$ to be present at the same time. However, it is perfectly possible to have both $Commit(\lceil\varphi\rceil)$ and $Commit(not\lceil\varphi\rceil)$ present in the commitment context at the same time. Due to the existence of schema K in this context, one can derive $Commit(\lceil\varphi\rceil \text{ and } not\lceil\varphi\rceil)$ which is different from $Commit(false)$. The argument is just a term and could have any semantics we wish to assign to it.

The beauty of multi-context systems is that it allows us to embed one logic as terms into another logic (the logic of the component). Therefore, $Commit(and(\varphi, not\ \varphi))$ evolves to $Commit(false)$ only if we apply propositional logic to the language modelled in this context. The two introspection axioms do not apply, since it does not make sense to say that once an agent is committed to some cause, it is also committed to be committed to this cause; similarly for negative introspection. All units also have modus ponens, uniform substitution and all tautologies from propositional logic.

4.1 Bridge Rules

There are a number of relationships between contexts that are captured by so-called bridge rules. A bridge rule of the form:

$$\frac{u_1 : \varphi, u_2 : \psi}{u_3 : \theta}$$

or sometimes written on one line for convenience as follows:

$$u_1 : \varphi, u_2 : \psi \rightarrow u_3 : \theta$$

can be read as: if the formula φ can be deduced in context u_1 and ψ in u_2 then the formula θ is to be added to the theory of context u_3 . It allows to relate formulae in one context to those in another. In [25] three different sets of bridge rules are described which model realistic, strongly realistic and weakly realistic-minded agents.

Figure 1 shows the model of an strongly realistic agent. Note, that in this figure, the C stands for a standard *communication* context as the authors did not concern themselves with commitments. One of its bridge rules, for example, states that something that is not desired should also not be intended.

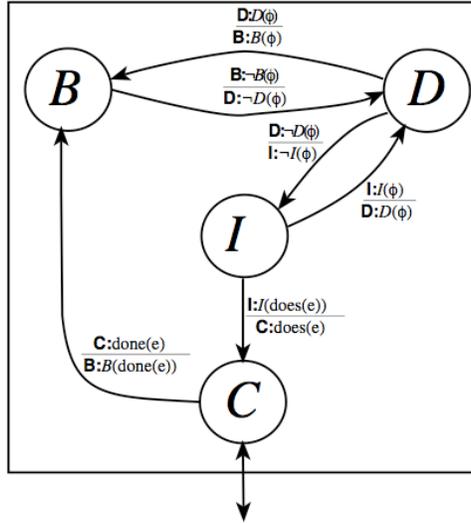


Figure 1: Multi-context description of a strongly realist BDI agent taken from [25].

In addition to the information-propagating bridge rules in the figure, there are more complex rules related to awareness of intention and impulsiveness between the belief and intention units (see [24]). These are common to all strongly realistic agents. Finally, there are domain specific rules, which link the contexts to the communication unit and control the impact of interaction with the environment on the mental state of an agent. An example of this is the bridge rule that stipulates that everything that is communicated to an agent to be done is believed to be done.

We are proposing to add an extra layer of bridge rules to existing BDI multi-context agents that controls the content of the mental contexts via norms. Remember, we earlier stated, that an adopted norm becomes a conditional commitment. The default *normative* personality of an agent is expressed as follows:

- an agent commits to believe everything it believes, commits to desire everything it desires and commits to intend everything it intends
- an agent believes what it is committed to believing, desires what it is committed to desiring and intends what it is committed to intending

This is modelled via two sets of bridge rules where Φ stands for any of B, D or I:

$$\frac{\Phi : \Phi(\varphi)}{CC : Commit([\Phi(\varphi)])} (*) \qquad \frac{CC : Commit([\Phi(\varphi)])}{\Phi : \Phi(\varphi)} (**)$$

Two examples of this can be seen in Figure 2 that depicts the normative layer of bridge rules we propose. Here CC represents the commitment context in order to distinguish it from the communication context C in Figure 1. Formulae (for now restricted to mental literals) from any of the three standard contexts are injected into the commitment context via a bridge rule of the form (*), where they encounter norms (first-order logic implications). Since the commitment context is closed under implication, the deduction machinery inside this context can be thought of as *applying* the norms. The resulting formulae of the local reasoning in the commitment unit are then injected back into the appropriate context via a bridge rule of the form (**).

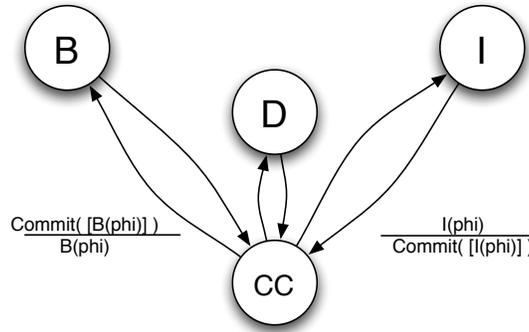


Figure 2: Commitment overlay for normative agents. In this figure, square brackets represent Goedel codification.

The six arcs in the figure represent the *default* normative personality of an agent. It is perfectly reasonable to imagine agents with different attitudes towards norms. A rebellious agent for example, may not desire or intend everything that it is committed to desiring or intending. Modelling agent types can therefore proceed on two levels. At the standard level between the belief-, desire- and intention context, personality traits like strong realism can be modelled, whereas character traits related to norms and norms adoption can be mimicked by modifying the overlay net of bridge rules involving the commitment context.

This proposed architecture is operationally speaking very simple. The complexity of norm execution is dealt with in the commitment context, whose logic is easily modifiable. Our modular, layered approach is a natural, clean extension that provides BDI agents with a new capability, namely norm compliance.

5 Truth Maintenance Problem

Adopting a norm and hence adding a commitment is in some way like opening a channel, linking different parts in the agent's 'brain'. For example, a commitment of the form $Commit(\lceil B(\varphi) \rightarrow I(\psi) \rceil)$ causes $I(\psi)$ to be deduced in the intention context if $B(\varphi)$ is deducible in the belief context. The reasoning is as follows:

- a bridge rule from the normative layer (see (*) in Section 4) adds $Commit(\lceil B(\varphi) \rceil)$ to the commitment context since $B(\varphi)$ is deducible in the belief context
- the adopted norm together with an instance of schema K allow us to deduce $Commit(\lceil I(\psi) \rceil)$
- another normative bridge rule (see (**)) injects $I(\psi)$ into the intention context

One can therefore think of this as having a bridge rule of the form:

$$\frac{B : B(\varphi)}{I : I(\psi)}$$

which is only activated, once $Commit(\lceil B(\varphi) \rightarrow I(\psi) \rceil)$ is present in the commitment context. What happens however, if $B(\varphi)$ is removed from the belief context? Should one also remove $I(\psi)$ from the intention context? What impact does the revocation of the commitment have?

In any case, one has to ensure the consistency of all the mental attitudes (since their respective logics contain schema D). Generally, *adopting a norm* has extensive ramifications. Firstly, a new commitment is added to the commitment context, possibly resulting in the local deduction of more commitments. The set of commitments needs to be maintained in a consistent state. Secondly, using the new commitment, contexts that feature in the norm description are linked. This may subsequently lead to the deduction of new formulae in these linked contexts. However, these newly deduced formulae can also be inconsistent with the existing theory of the context in question. This dilemma is known as the *truth maintenance (TM) problem*. Artificial Intelligence has seen many different approaches to the TM problem in the past. We believe, for our purposes, the most promising ones are:

5.1 Standard truth maintenance systems

Once a bridge rule is fired and tries to inject a formula into a context, it is the responsibility of the context to maintain consistency. In the simplest case, it checks, whether

the formula to be inserted is inconsistent with the existing theory of the context and if it finds this to be the case, rejects the proposed injection. If no conflict is detected (i.e. the opposite of the formula to be injected is not present in the current theory of the context), then the formula is added to the theory.

This is a very simplistic approach that only allows monotonic updates. More sophisticated truth maintenance systems can handle non-monotonic updates or *belief revision*. A formula which contradicts the existing theory in a context can still be inserted, but some machinery must then revise the theory to make it consistent again (by removing some of the causes of the contradiction). Two main approaches whose use we are investigating currently are *justification-based* truth maintenance systems like the one proposed by Doyle [12] and *assumption-based* truth maintenance systems following the work by deKleer [9].

In particular, we are interested in two different kinds of inconsistencies. One of them is the inconsistency inside a context, caused by the injection of formulae that contradict the existing theory. This problem, we intend to address using a variant of either Doyle's or deKleer's TMS. The second type of inconsistency comes about by adopting conflicting commitments. For example, a norm that prohibits more than two consecutive dances with the same partner is in conflict with a norm that states that an experienced dancer should dance three consecutive dances with a beginner, if requested. We are currently working with García-Camino on resolving conflicts between norms.

5.2 Argumentation

A traditional way of resolving conflicts is to consider the arguments in favour and against a decision and choose those that are more convincing. The area of *argumentation* studies this process and gives tools, mostly based on logical approaches, to automate this decision process (see for example the work detailed in [3] and [25]). In these works the decision is made considering that arguments are proofs in a logical formalism and that the proofs *attack* one another by deducing opposite literals or *rebut* one another by deducing the opposite of a literal used to support a proof.

In our case we need to have a notion of argument that bases the *attack* not *only* on some logical relationships between the proofs used to support two opposite literals, but also the fact that some of the proofs are based on the application of norms. Therefore, beyond the logical attack we'll have to consider the *strength* of the argument in how supported it is by the norms of the particular institution. In that sense, we suggest to include in the proof the set of norms applied to generate a commitment and use a measure over them when the content of the commitment challenges a pre-existing intention, belief or desire. This measure can be based on specific reasoning with respect to the willingness that the agent has to respect a given norm, or the *degree of adoption* of the norm by the agent.

5.3 Decision Theory and Graded Mental Attitudes

Decision theory on the other hand is based on *utilities*. It decrees that decisions have to be made based on the most valuable outcome. When faced with two conflicting intentions, one dictated by a norm and the opposite dictated by the agent’s desires, it may decide to violate the norm, if this violation and the fulfilment of its desire are more *satisfying* than conforming to the norm. In order to decide what is more satisfying, we propose to use graded mental attitudes similar to the work done by Casali et al. [4].

In their work, the atomic formulae in contexts are no longer of the form $B(term)$ but instead are enriched with a weight ϵ to give $B(term, \epsilon)$. This weight represents the degree of belief. Similarly, for desires, it represents the degree of desire allowing us to attach priorities to certain formulae. In the case of intentions, the weight can be used to model the cost/benefit trade-off of the currently intended action. Finally, a weight on a commitment indicates the degree of adoption. Using these graded modalities, one can compute the utility of each of the conflicting atoms and act accordingly.

6 Towards Norm Adoption

In this section, we will demonstrate how our architecture can be used to model norm adoption and related issues. In Figure 3 one can see a dancer agent consisting of the four contexts described before. In addition, the agent has a communication context, denoted C, that is connected to some of the other contexts (e.g. in Figure 1 it is connected only to the belief context).

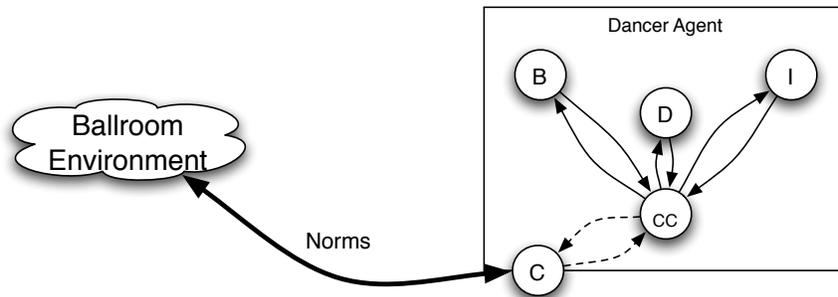


Figure 3: Agent adopting norms from the environment

Here we concern ourselves with the relationship between the commitment and communication contexts. The communication context is the agent’s connection with the environment and is used (among other things) to learn about new norms and communicate new norms to other agents.

The nature of the bridge rules between these two contexts determine the attitude an agent takes towards norm adoption. An obedient agent, for example, would adopt any norm that it is subjected to, regardless of potential conflicts this would cause with respect to its beliefs, desires and intentions. In the ballroom setting, the environment could announce that the remainder of the evening is spend as a *ladies' choice ball* by communicating the norm $\mathbb{B}(\text{sex}(X, \text{male})) \rightarrow \neg\mathbb{I}(\text{initiateConversation}(X, _))$. A male dancer who currently has the intention to invite a female dancer is now facing a decision. If he adopts the norm, he will be unable to fulfil his intention. However, rejecting the norm (or accepting and violating it) may have social repercussions. We are currently investigating the issue of norm adoption in our BDI+C framework and expect that it is conducive to a formal analysis of this topic.

7 Related Work

We have referred to related work throughout the paper, in particular in the first half. In this section, we aim to contrast our proposal with two particular lines of work, namely a modified BDI interpreter by Dignum et. al. [10] and the BOID architecture by Broersen et. al. [2]. Dignum and his colleagues add one step to the main loop of the BDI interpreter in which selected events are augmented with deontic events by repeatedly applying the introspective norms and obligations [10]. They distinguish between norms (that hold for a society) and obligations (that hold between two agents). They rank obligations based on the punishments associated with their violation and norms based on their social benefit. Our view of commitments is broader in that we allow the committed entities and the subjects of the commitment to be agents, groups of agents or entire societies. The architecture we propose is more flexible, too, since each component has its own logic and the relationships between components can be varied dynamically.

The BOID architecture by Broersen and his colleagues has many similarities with our work. It contains four components (B, O, I and D) where the O component stands for obligations (as opposed to commitments in our case) and the other components have the usual meaning. They suggest feedback loops that feed the output of every component back to the belief component for reconsideration. The order in which components are chosen for rule selection, determines the kind of character the agent possesses. For example, if obligations are considered before desires, the agent is deemed to be social. One drawback is, that they only consider orders in which the belief component overrules any other modality [2]. Furthermore, these orders are fixed for each agent. Using our agent architecture, agent types can be modified dynamically. Also, the relationship between mental attitudes can be controlled at a finer level of granularity (e.g. domain-specific rules connecting multiple contexts rather than the strict ordering of components required in the BOID architecture).

8 Future Work

We are currently working on implementing the BDI+C agent architecture using the programming language QuP++ [6] an object-oriented extension of QuProlog [27]. The advantage of this particular Prolog variant is its multi-threadedness and support for reasoning. We are implementing every context as an individual thread and use separate threads for bridge rules to synchronise between the contexts. Another line of research is concerned with generalising both the architecture and the implementation to handle graded mental attitudes. Casali et al. [4] have formalised the notion of uncertainty for the BDI model and we believe it can be employed in the BDI+C model in order to tackle the truth maintenance problem. Furthermore, it will allow us to represent the character or type of an agent more closely. We even envisage the ability to express the *mood* of an agent via dynamically changing the degree to which it believes, desires, intends and sticks to its commitments.

Furthermore, our interest lies in investigating temporal aspects of norms and norm adoption. In [28], Sabater et al. extend the syntax of bridge rules by introducing the notions of consumption and time-outs. We intend to make use of these extensions in order to allow for more expressiveness in the formulation of normative commitments.

López y López, in her doctoral thesis [19], describes different strategies for norm adoption ranging from fearful, rebellious and greedy character traits to reciprocation and imitation of other agents. All her strategies are based on potential rewards or punishments. Broersen et. al. in [2] define agent characters based on the fixed order of the belief-, obligation-, intention- and desire-component (though they do not use multi-contexts, one can think of their components as such). They also give names such as ‘super-selfish’ to some of these orderings. Using the extended bridge rules of our architecture combined with graded versions of the mental attitudes, we can define different agent characters more formally and on a much finer level of granularity. The personality type indicator developed by Katharine Cook Briggs and her daughter (see also the MBTI manual [21]) classifies individuals along four dimensions. For example, they differentiate between extroverted and introverted individuals and distinguish those that are very logical from those that are lead by intuition. We hope to apply some of their classifiers to our agent modelling.

The ballroom case study provides further opportunities to experiment with. For example, more than two dancers can be involved in some kind of square dance or minuet. The notions of release from commitment and norm evolution are also very interesting in this context. We intend to stretch the applicability of our proposed agent architecture to find out its limitations and possibly expand it.

9 Conclusions

In this paper we have outlined a conservative extension of BDI agent architectures to grasp the notion of commitments and we have further shown how to use this extension to express norm adoption by such agents. We have proposed how to make these extensions operational in terms of multi-context logics and illustrated them with an example of dance negotiations following the etiquette conventions of a ballroom.

We found that our proposed extension of a BDI architecture—to incorporate the notion of commitment—has the following features:

1. It is easy to describe, formalise and make operational.
2. It may be readily added on top of a given BDI model by simply including a new context and bridge rule schemata linking it to each of the other modalities.
3. Although we have proposed a schema that is uniform for all modalities, it is easy to tune-tune any given formalisation of the features of the commitment unit and the underlying BDI architecture in order to capture alternative formalisations, shades of meaning or the character or personality of an agent.
4. Our BDI+C model appears to be general enough to explore with it the complex aspects of legal consequence; especially in its concrete aspects of individual norm compliance with respect to the attitude of an agent towards authority, utility, selfishness and other features that have been addressed by the MAS community.
5. The notion of norms as an initial theory for the commitment context and the commitment-dependent bridge rules provide convenient ways to study para-normative aspects like norm adoption, compliance, blame assignment, violation, reparation or hierarchical normative sources. Likewise, the notion of contract could be modelled as joint commitments and added to the commitment context.
6. In a similar fashion, we have only pointed out a straightforward translation of norms as commitments between individuals and an institution, although it should be evident that other notions of authority (hierarchies of norms, issuers of norms, contingent applicabilities of norms) may be modelled along the lines we outline in this paper.
7. The evolution of the belief-, desire-, intention- and commitment theories as interaction proceeds and associated consistency issues may be addressed with the type of tools that have been applied to other dynamic theories, although in this paper we only hinted at three mechanisms: standard truth-maintenance systems, graded versions of the modalities and argumentation.

8. The modularity of the architecture and its simple operationalisation suggest the possibility of implementations that may have interesting experimental and simulation properties to study the underlying theoretical and formal aspects as well as the applications of this architecture in multi-agent systems.

Acknowledgements

The authors would like to thank Keith Clark for his valuable comments and the Spanish Ministry of Education and Science (MEC) for support through the Web-i-2 project (TIC-2003-08763-C02-00). The first author is also grateful for a student grant from the Spanish Scientific Research Council through the Web-i(2) project (CSIC PI 2004-5 0E 133).

References

- [1] Josep Ll. Arcos, Marc Esteva, Pablo Noriega, Juan Antonio Rodríguez, and Carles Sierra. Environment engineering for multiagent systems. *Journal on Engineering Applications of Artificial Intelligence*, 18(2):191–204, 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: Proceedings of the fifth international conference on Autonomous agents*, pages 9–16, New York, NY, USA, 2001. ACM Press.
- [3] Marcela Capobianco, Carlos I. Chesñevar, and Guillermo Simari. Argumentation and the dynamics of warranted beliefs in changing environments. *Intl. Journal on Autonomous Agents and Multiagent Systems (JAAMAS)*, 11:127–151, September 2005.
- [4] Ana Casali, Lluís Godo, and Carles Sierra. Graded BDI models for agent architectures. In *5th International Workshop on Computational Logic in Multi-Agent Systems (CLIMA V)*, pages 126–143, Lisbon, Portugal, 2004.
- [5] Cristiano Castelfranchi. Commitment: from intentions to groups and organizations. In *Proceedings of ICMAS'95*, pages 41–48, Cambridge (MA), 1995. AAAI/MIT Press.
- [6] Keith L. Clark and Peter J Robinson. *Computational logic: logic programming and beyond: essays in honour of Robert A. Kowalski, Part I*, chapter Agents as multi-threaded logical objects, pages 33–65. Springer, 2002.
- [7] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213–261, 1990.

- [8] Philip R. Cohen and Hector J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communications*, pages 221–255. MIT Press, 1990.
- [9] Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.
- [10] F. Dignum, D. Morley, E. Sonenberg, and L. Cavendon. Towards socially sophisticated BDI agents. In E. H. Durfee, editor, *ICMAS '00: Proceedings of the Fourth International Conference on MultiAgent Systems (ICMAS-2000)*, pages 111–118, Washington, DC, USA, 2000. IEEE Computer Society.
- [11] Mark d’Inverno, David Kinny, Michael Luck, and Michael Wooldridge. A formal specification of dMARS. In *Agent Theories, Architectures, and Languages*, pages 155–176, 1997.
- [12] Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979.
- [13] Dorian Gaertner, Keith L. Clark, and Marek J. Sergot. Ballroom etiquette: a case study for norm-governed multi-agent systems. In *Proceedings of the First International Workshop on Coordination, Organization, Institutions and Norms (COIN) 2006 at AAMAS, Hakodate, Japan, 2006*.
- [14] Andrés García-Camino, Pablo Noriega, and Juan Antonio Rodríguez-Aguilar. Implementing norms in electronic institutions. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (AAMAS 2005)*, pages 667–673, New York, NY, USA, 2005. ACM Press.
- [15] Andrés García-Camino, Juan Antonio Rodríguez-Aguilar, Carles Sierra, and Wamberto Vasconcelos. A distributed architecture for norm-aware agent societies. In *Fourth International Joint Conference on Autonomous Agents and Multiagent Systems. Declarative Agent Languages and Technologies workshop (DALT’05)*, 2005.
- [16] Fausto Giunchiglia and Luciano Serafini. Multilanguage hierarchical logics or: How we can do without modal logics. *Artificial Intelligence*, 65(1):29–70, 1994.
- [17] John Knottenbelt and Keith L. Clark. Contract related agents. In *Sixth International Workshop on Computational Logic in Multi-Agent Systems (CLIMA VI)*, 2005.
- [18] Robert A. Kowalski and Marek J. Sergot. A logic-based calculus of events. *New Gen. Comput.*, 4(1):67–95, 1986.
- [19] Fabiola López y Lopéz. *Social Powers and Norms: Impact on Agent Behaviour*. PhD thesis, University of Southampton, 2003.
- [20] Fabiola López y Lopéz, Michael Luck, and Mark d’Inverno. A normative framework for agent-based systems. In *Proceedings of the 1st International Symposium on Normative Multiagent Systems*, 2005.

- [21] Isabel Briggs Myers, Mary H. McCaulley, Naomi L. Quenk, and Allen L. Hammer. *MBTI Manual (A guide to the development and use of the Myers Briggs type indicator)*. Consulting Psychologists Press, 3rd edition, 1998.
- [22] Pablo Noriega and Carles Sierra. Towards layered dialogical agents. In *ECAI '96: Proceedings of the Workshop on Intelligent Agents III, Agent Theories, Architectures, and Languages*, pages 173–188, London, UK, 1997. Springer-Verlag.
- [23] Timothy J. Norman and Nicholas R. Jennings. Generating states of joint commitment between autonomous agents. *Lecture Notes in Computer Science*, 1441:123–134, 1998.
- [24] Simon Parsons, Nicholas R. Jennings, Jordi Sabater, and Carles Sierra. Agent specification using multi-context systems. In *Selected papers from the UKMAS Workshop on Foundations and Applications of Multi-Agent Systems*, pages 205–226, London, UK, 2002. Springer-Verlag.
- [25] Simon Parsons, Carles Sierra, and Nick Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261–292, 1998.
- [26] Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multiagent Systems (ICMAS)*, pages 312–319, San Francisco, California, USA, 1995.
- [27] Peter J. Robinson, Mike G. Hinchey, and Keith L. Clark. Qu-Prolog: an implementation language for agents with advanced reasoning capabilities. In M G Hinchey, J L Rash, W E Truszkowski, C Rouff, and D GordonSpears, editors, *2nd international workshop on formal approaches to agent-based systems (FAABS 2002)*, pages 162–172, Greenbelt, Maryland, USA, 2003. Springer-Verlag.
- [28] Jordi Sabater, Carles Sierra, Simon Parsons, and Nicholas R. Jennings. Engineering executable agents using multi-context systems. *J. Log. Comput.*, 12(3):413–442, 2002.