A Normative Framework for Agent-Based Systems

Fabiola López y López[⋆]
[⋆]University of Puebla
México
fabiola@cs.buap.mx

Michael Luck[†]
[†]University of Southampton
United Kingdom
mml@ecs.soton.ac.uk

Mark d'Inverno[‡]
[‡]University of Westminster
United Kingdom
dinverm@westminster.ac.uk

Abstract

One of the key issues in the computational representation of *open societies* relates to the introduction of *norms* that help to cope with the heterogeneity, the autonomy and the diversity of interests among their members. Research regarding this issue presents two omissions. One is the lack of a canonical model of norms that facilitates their implementation, and that allows us to describe the processes of reasoning about norms. The other refers to considering, in the model of normative multi-agent systems, the perspective of individual agents and what they might need to effectively reason about the society in which they participate. Both are the concerns of this paper, and the main objective is to present a formal normative framework for agent-based systems.

1 Introduction

Norms have long been used as mechanisms to limit human autonomy in such a way that coexistence between self-interested and untrusted people has been made possible. They are indispensable to overcome problems of coordination of large, complex and heterogeneous systems where total and direct social control cannot be exerted. From this experience, the introduction of *norms* that help to cope with the heterogeneity, the autonomy and the diversity of interests among agents has been considered as a key issue towards the computational representation of *open societies* of agents (Luck et al., 2003).

Although efforts have been made to describe and define the different types of norms that agents have to deal with (Dignum, 1999; Singh, 1999), work has not led into a model that facilitates the computational representation of any kind of norm. Each kind of norm appears to be different, which also suggests that different processes of reasoning should be proposed. There are some work that introduces norms in systems of agents to represent societies, institutions and organisations (Dellarocas and Klein, 2001; Dignum and Dignum, 2001; Esteva et al., 2001; Shoham and Tennenholtz, 1995). This research is primarily focused at the level of multi-agent systems, where norms represent the means to achieve coordination among their members. There, agents are assumed to be able to comply with norms, to adopt new norms, and to obey the authorities of the system but nothing is said about the reasons why agents might be willing to adopt and comply with norms, nor about how agents can identify situations in which an authority's orders are beyond its responsibilities. That is, although agents in such systems are said to be autonomous, their models of norms and systems regulated by norms do not offer the means to explain why *autonomous* agents that are working to satisfy their own goals, still comply with their social responsibilities. In addition, although the importance of modelling compliance with norms as an autonomous decision has been identified by several researchers (Castelfranchi et al., 2000; Conte et al., 1999a; Conte and Dellarocas, 2001; Conte et al., 1999), the issue is only partly addressed by others whose proposals for norm compliance generally rely on specific decision-making strategies based on how much an agent gains or loses by complying with (Barbuceanu et al., 1999; Dignum et al., 2000), and on the probability of being caught by a defender of a norm (Boella and Lesmo, 2001). We consider these cases as very specific and, therefore, inadequate to model different kinds of normative behaviour of autonomous agents.

As a way to overcome these omissions, we have developed a normative framework for agent-based systems that includes a canonical model of norms, a model of normative multi-agent systems and a model of normative autonomous agents. Independent components of this framework have already been presented in different forums (López and Luck, 2003, 2004; López et al., 2002, 2004). The objective of this paper is to present the framework as a whole. The formal model presented in this paper is written in the Z language, which is based on set-theory and

first order logic (Spivey, 1992). The organisation of the paper is as follows. First, a formal definition of an autonomous agents is given. After that, an analysis of different properties of norms is provided. This analysis is then used to justify the elements that a general model of a norm must include in order to enable autonomous agents to reason about them. Next, the main properties of systems of autonomous agents that are regulated by norms are discussed and a model is presented. Then, we describe our proposal to enable agents to reason about norms. Finally, our conclusions are provided.

2 Autonomous Agents

The foundations of this work are taken from Luck and d'Inverno's SMART agent framework (d'Inverno and Luck, 2003) whose concept of *motivations* as the driving force that affects the reasoning of agents in satisfying their goals is considered as the underlying argument for agents to voluntarily comply with norms and to voluntarily enter and remain in a society. In the SMART agent framework, an *attribute* represents a perceivable feature of the agent's environment, which can be represented as a predicate or its negation. Then, a particular *state* in the environment is described by a set of attributes, a *goal* represents situations that an agent wishes to bring about, *motivations* are desires or preferences that affect the outcome of the reasoning intended to satisfy an agent's goals, and *actions* are discrete events that change the state of the environment when performed. For the purposes of this paper, we formally describe environmental states, goals, actions and autonomous agents. Details of the remaining elements are not needed, so we simply consider them as given sets.

$$[Attribute, Motivation]$$

$$EnvState == \mathbb{P}_1\, Attribute$$
$$Goal == \mathbb{P}_1\, Attribute$$
$$Action == EnvState \rightarrow EnvState$$

AutonomousAgent

$$goals : \mathbb{P}\, Goal; \quad capabilities : \mathbb{P}\, Action;$$
$$motivations : \mathbb{P}\, Motivation;$$
$$beliefs : \mathbb{P}_1\, Attribute$$
$$importance : \mathbb{P}(\mathbb{P}\, Goal \times \mathbb{P}\, Motivation) \rightarrow \mathbb{N}$$

$$goals \neq \varnothing; \qquad motivations \neq \varnothing$$
$$\forall\, x : \mathbb{P}\, Goal, y : \mathbb{P}\, Motivation \bullet$$
$$(x, y) \in \mathrm{dom}\, importance \mid$$
$$x \subseteq goals \land y \subseteq motivations$$

In the above schema, an autonomous agent is described by a set of goals that it wants to bring about, a set of capabilities that it is able to perform, a non-empty set of motivations representing its preferences, and a set of beliefs representing its vision about the external world. We also assume that the agent is able to determine the $importance$ of its goals, which depends on its current motivations.

3 Norms

Norms facilitate mechanisms to drive the behaviour of agents, especially in those cases when their behaviour affects other agents. Norms can be characterised by their *prescriptiveness*, *sociality*, and *social pressure*. In other words,

- a norm tells an agent how to behave (*prescriptiveness*);
- in situations where more than one agent is involved (*sociality*);
- and since it is always expected that norms conflict with the personal interest of some agents, socially acceptable mechanisms to force agents to comply with norms are needed (*social pressure*).

By analysing these properties, the essential components of a norm can be identified.

3.1 Norm Components

Norms specify patterns of behaviour for a set of agents. These patterns are sometimes represented as actions to be performed (Axelrod, 1986; Tuomela, 1995), or restrictions to be imposed over an agent's actions (Norman et al., 1998; Shoham and Tennenholtz, 1995). At other times, patterns of behaviour are specified through goals that must either be satisfied or avoided by agents (Conte and Castelfranchi, 1995; Singh, 1999). Now, since actions are performed in order to change the state of an environment, goals are states that agents want to bring about, and restrictions can be seen as goals to be avoided, we argue that by considering goals the other two patterns of behaviour can be easily represented (as shown in (López and Luck, 2003)).

In brief, norms specify things that ought to be done and, consequently, a set of *normative goals* must be included. Sometimes, these normative goals must be directly intended, while at other times their role is to inhibit specific states (as in the case of prohibitions). Norms are always directed at a set of *addressee agents*, which are directly responsible for the satisfaction of the normative goals. Moreover, sometimes to take decisions regarding norms, agents not

only consider what must be done but also for whom it must be done. Then, agents that *benefit* from the satisfaction of normative goals may also be included.

In general, norms are not applied all the time, but only in particular circumstances or within a specific *context*. Thus, norms must always specify the situations in which addressee agents must fulfill them. *Exception* states may also be included to represent situations in which addressees cannot be punished when they *have not* complied with norms. Exceptions represent *immunity* states for all addressee agents in a particular situation (Ross, 1968). Now, to ensure that personal interests do not impede the fulfillment of norms, mechanisms either to promote compliance with norms, or to inhibit deviation from them, are needed. Norms may include *rewards* to be given when normative goals become satisfied, or *punishments* to be applied when they are not. Both rewards and punishments are the means for addressee agents to determine what might happen whatever decision they take regarding norms. They are not the responsibility of addressees agents but of other agents already entitled to either reward or punish compliance and non-compliance with norms. Since rewards and punishments represent states to be achieved, it is natural to consider them as goals but, in contrast with normative goals that must be satisfied by addressees, punishments and rewards are satisfied by agents entitled to do so.

In other words, a norm must be considered for fulfillment by an agent when certain environmental states, not included as exception states, hold. Such a norm forces a group of addressee agents to satisfy some normative goals for a (possibly empty) set of beneficiary agents. In addition, agents are aware that rewards may be enjoyed if norms become satisfied, or that punishments that affect their current goals can be applied if not. The formal specification of a norm is given in the $Norm$ schema where all the components of norms described here are included, together with some constraints on them. First, it does not make any sense to have norms specifying nothing, norms directed at nobody, or norms that either never or always become applied. Thus, the first three predicates in the schema state that the set of normative goals, the set of addressee agents, and the context must never be empty. The fourth predicate states that the set of attributes describing both the context and exceptions must be disjoint to avoid inconsistencies in identifying whether a norm must be applied. The final constraint specifies that punishments and rewards are also consistent and, therefore, they must be disjoint.

$$
\begin{array}{|l}
\hline
Norm \\
\hline
normativegoals : \mathbb{P}\ Goal \\
addressees : \mathbb{P}\ NormativeAgent \\
beneficiaries : \mathbb{P}\ NormativeAgent \\
context : EnvState \\
exceptions : EnvState \\
rewards : \mathbb{P}\ Goal \\
punishments : \mathbb{P}\ Goal \\
\hline
normativegoals \neq \varnothing \\
addressees \neq \varnothing \\
context \neq \varnothing \\
context \cap exceptions = \varnothing \\
rewards \cap punishments = \varnothing \\
\hline
\end{array}
$$

3.2 Considerations

The term *norm* has been used as a synonym for obligations (Boella and Lesmo, 2001; Dignum et al., 2000), prohibitions (Dignum, 1999), social laws (Shoham and Tennenholtz, 1995), and other kinds of rules imposed by societies (or by an authority). The position of our work is quite different. It considers that all these terms can be grouped in a general definition of a norm, because they have the same properties (i.e. prescriptiveness, sociality and social pressure) and they can be represented by the same model. They all represent responsibilities for addressee agents, and create expectations for beneficiaries and other agents. They are also the means to support beneficiaries when they have to claim some compensation in the situations where norms are not fulfilled as expected. Moreover, whatever the kind of norm being considered, its fulfillment may be rewarded, and its violation may be penalised. What makes one norm different from another is the way in which they are created, their persistence, and the components that are obligatory in the norm. Thus, norms might be created by an agent designer as built-in norms, they can be the result of agreements between agents, or they can be elaborated by a complex legal system. Regarding their persistence, norms might be taken into account during different periods of time, such as until an agent dies, as long as an agent stays in a society, or just for a short period of time until its normative goals become satisfied. Finally, some components of a norm might not exist; there are norms that include neither punishments nor rewards, even though they are complied with. Some of these characteristics can be used to provide a *classification* of norms into four main categories: *obligations*, *prohibitions*, *social commitments* and *social codes*. Despite these differences, all types of norms can be reasoned about in similar ways.

Now, to understand the consequences of norms in a particular system, it is necessary to consider norms

that are either *fulfilled* or *unfulfilled*. However, since most of the time a norm has a set of agents as addressees, the meaning of fulfilling a norm might depend on the interpretation of analysers of a system. In small groups of agents, it might be easy to consider a norm as fulfilled when every addressee agent has fulfilled the norm; by contrast, in larger societies, a proportion of agents complying with a norm will be enough to consider it as fulfilled. Instead of defining fulfilled norms in general, it is more appropriate to define norms being fulfilled by a particular addressee agent. To do so, the concept of norm instances is introduced as follows. Once a norm is adopted by an agent, a *norm instance* is created, which represents the internalisation of a norm by an agent (Conte and Castelfranchi, 1995). A norm instance is a copy of the original norm that is now used as a *mental attitude* from which new goals for the agent might be inferred. Norms and norm instances are the same concept used for different purposes. Norms are abstract specifications that exist in a society and are known by all agents (Tuomela, 1995), but agents work with *instances* of these norms. Consequently, there must be a separate instance for each addressee of a norm. Due to space constraints, formal definitions and examples of categories of norms, norm instances and fulfilled norms are not provided here but can be found elsewhere (López and Luck, 2003).

3.3 Interlocking Norms

The norms of a system are not isolated from each other; sometimes, compliance with them is a condition to trigger (or activate) other norms. That is, there are norms that prescribe how some agents must behave in situations in which other agents either comply with a norm or do not comply with it (Ross, 1968). For example, when employees comply with their obligations in an office, paying their salary becomes an obligation of the employer; or when a plane cannot take-off, providing accommodation to passengers becomes a responsibility of the airline. Norms related in this way can make a complete chain of norms because the newly activated norms can, in turn, activate new ones. Now, since triggering a norm depends on past compliance with another norm, we call these kinds of norms *interlocking norms*. The norm that gives rise to another norm is called the *primary* norm, whereas the norm activated as a result of either the fulfillment or violation of the first is called the *secondary* norm.

In terms of the norm model mentioned earlier, the *context* is a state that must hold for a norm to be complied with. Since the fulfillment of a norm is assessed through its normative goals, the context of the secondary norm must include the satisfaction (or non-satisfaction) of all the primary norm's normative goals. Figure 1 illustrates the structure of both the primary and the secondary norms and how they are interlocked through the primary norm's normative goals and the secondary norm's context.

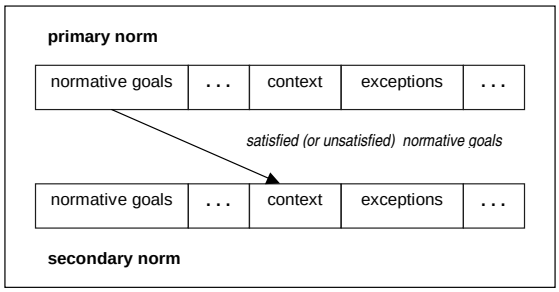

Figure 1: Interlocking Norm Structure

Formally, a norm is interlocked with another norm *by non-compliance* if, in the context of the secondary norm, an instance of the primary norm can be considered as violated. This means that when any addressee of a norm does not fulfill the norm, the corresponding interlocking norm will be triggered. The formal specification of this is given below, where n_1 represents the primary norm and n_2 is the secondary norm.

$$
\begin{array}{|l}
lockedbynoncompliance\_ : \mathbb{P}(Norm \times Norm) \\
\hline
\forall\, n_1, n_2 : Norm \bullet \\
\quad lockedbynoncompliance\ (n_1, n_2) \Leftrightarrow \\
\quad\quad (\exists\, ni : NormInstance \mid \\
\quad\quad isnorminstance\ (ni, n_1) \bullet \\
\quad\quad\quad \neg\ fulfilled\ (ni, n_2.context))
\end{array}
$$

Similarly, a norm is interlocked with another norm *by compliance* if, in the context of the secondary norm, an instance of the primary norm can be considered as fulfilled. Thus, any addressee of the norm that fulfills it will trigger the interlocking norm. The specification of this is given as follows.

$$
\begin{array}{|l}
lockedbycompliance\_ : \mathbb{P}(Norm \times Norm) \\
\hline
\forall\, n_1, n_2 : Norm \bullet \\
\quad lockedbycompliance\ (n_1, n_2) \Leftrightarrow \\
\quad\quad (\exists\, ni : NormInstance \mid \\
\quad\quad isnorminstance\ (ni, n_1) \bullet \\
\quad\quad\quad fulfilled\ (ni, n_2.context))
\end{array}
$$

Having the means to relate norms in this way allows us to model how the normative behaviour of agents that are addressees of a secondary norm is *influenced* by the normative behaviour of addressees of a primary norm.

4 Normative Multi-Agent Systems

Since norms are social concepts, they cannot be studied independently of the systems for which they are created and, consequently, an analysis of the normative aspects of social systems must be provided. Although social systems that are regulated by norms are different from one another, some general characteristics can be identified. They consist of a set of agents that are controlled by the same set of norms ranging from obligations and social commitments to social codes. However, whereas there are static systems in which all norms are defined in advance and agents in the system always comply with them (Boman, 1999; Shoham and Tennenholtz, 1995), a more realistic view of these kinds of systems suggests that when *autonomous* agents are considered, neither can all norms be known in advance (since new conflicts among agents may emerge and, therefore, new norms may be needed), nor can compliance with norms be guaranteed (since agents can decide not to comply). We can say then, that systems regulated by norms must include mechanisms to deal with both the modification of norms and the unpredictable normative behaviour of autonomous agents. So, *normative multi-agent systems* have the following characteristics.

- *Membership*. Agents in a society must be able to deal with norms but, above all, they must recognise themselves as part of the system. This kind of social identification means that agents adopt the society norms and, by doing so, they show their willingness to comply with these norms.
- *Social Pressure*. Effective authority cannot be exerted if penalties or incentives are not applied when norms are either violated or complied with. However, this control must not be an agent's arbitrary decision, and although it is only exerted by some agents, it must be socially accepted.
- *Dynamism*. Normative systems are *dynamic* by nature. New norms are created and obsolete norms are abolished. Compliance or non-compliance with norms may activate other norms and, therefore, force other agents to act. Agents can either join or leave the system. The normative behaviour of agent members might be unexpected, and it may influence the behaviour of other agents.

Given these characteristics, we argue that multi-agent systems must include mechanisms to defend norms, to allow their modification, and to identify authorities. Moreover, their members must be agents able to deal with norms. Each one of these concepts is discussed in detail and formalised in (López and Luck, 2004), here, we present just a summary of them.

4.1 Normative Agents

The effectiveness of every structure of control relies on the capabilities of its members to recognise and follow its norms. However, given that agents are autonomous, the fulfillment of norms can never be taken for granted (López et al., 2002). A *normative agent* is an agent whose behaviour is partly shaped by norms. They are able to deal with norms because they can represent, adopt, and comply with them. However, for autonomous agents, decisions to adopt or comply with norms are made on the basis of their own goals and motivations. That is, autonomous agents are not only able to *act on* norms but also they are able to *reason about* them. In what follows, all normative agents are considered as autonomous agents that have adopted some norms (*norms*) and, has decided which norms to comply with (*intended norms*) and which norms to reject (*rejected norms*). Although their normative behaviour is described in the next section, their representation is given now in the schema below.

$$
\begin{array}{|l}
\hline
NormativeAgent \\
\hline
AutonomousAgent \\
norms, intented, rejected : \mathbb{P}\, Norm \\
\hline
intented \subseteq norms \\
rejected \subseteq norms \\
\hline
\end{array}
$$

4.2 Enforcement and Reward Norms

Particularly interesting for this work are the norms triggered in order to punish offenders of other norms. We call them *enforcement norms* and their addressees are the *defenders* of a norm. These norms represent exerted social pressure because they specify not only who must apply the punishments, but also under which circumstances these punishments must be applied (Ross, 1968). That is, once the violation of a norm becomes identified by defenders, their duty is to start a process in which offender agents can be punished. For example, if there is an obligation to pay accommodation fees for all students in a university, there must also be a norm stating what hall managers must do when a student refuses to pay.

As can be seen, norms that enforce other norms are a special case of interlocking norms because besides being interlocked by non-compliance, the normative goals of the secondary norm must include every punishment of the primary norm. Figure 2 shows how

the structures of both norms are related. By modelling enforcement norms in this way, we cause an offender's punishments to be consistent with a defender's responsibilities. Addressees of an *enforced* norm (i.e. the primary norm) know what could happen if the norm is not complied with, and addressees of an *enforcement* norm (i.e. the secondary norm) know what must be done in order to punish the offenders of another norm. Enforcement norms allow the authority of defenders to be clearly constrained.

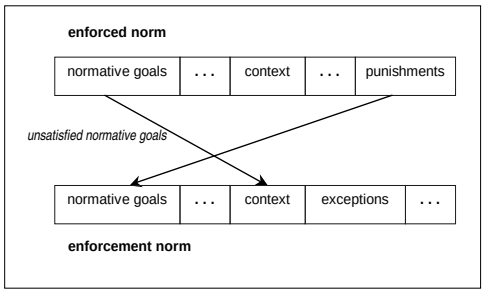

Figure 2: Enforcement Norm Structure

Formally, the relationship between a norm directed to control the behaviour of some agents and a norm directed at punishing the offenders of such a norm can be defined as follows. A norm *enforces* another norm if the first norm is activated when the second is violated, and all punishments associated with the violated norm are part of the normative goals of the first. Every norm satisfying this property is known as an *enforcement* norm.

$$enforces\_ : \mathbb{P}(Norm \times Norm)$$

$$\forall\, n_1, n_2 : Norm \bullet enforces\ (n_1, n_2) \Leftrightarrow lockedbynoncompliance\ (n_2, n_1) \wedge n_2.punishments \subseteq n_1.normativegoals$$

So far we have described some interlocking norms in terms of punishments because these are one of the more commonly used mechanisms to enforce compliance with norms. However, a similar analysis can be applied to interlocking norms corresponding to the process of rewarding members doing their duties. These norms must be interlocked by compliance and all the rewards included in the primary norm (rewarded norm) must be included in the normative goals of the secondary norm (reward norm). The relation between these norms is shown in Figure 3.

Formally, we say that a norm *encourages* compliance with another norm if the first norm is activated when the second norm becomes fulfilled, and the rewards associated with the fulfilled norm are part of

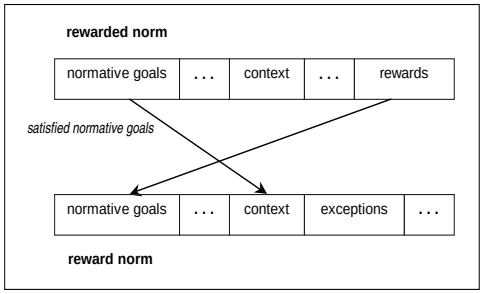

Figure 3: Reward Norm Structure

the normative goals of the first norm. Every norm satisfying this property is known as a *reward* norm.

$$rewardnorm\_ : \mathbb{P}(Norm \times Norm)$$

$$\forall\, n_1, n_2 : Norm \bullet rewardnorm\ (n_1, n_2) \Leftrightarrow lockedbycompliance\ (n_2, n_1) \wedge n_2.rewards \subseteq n_1.normativegoals$$

It is important to mention that this way of representing enforcement and reward norms can create an infinite chain of norms because we would also have to define norms to apply when authorities or defenders do not comply with their obligations, either to punish those agents breaking rules or to reward those agents that fulfill their responsibilities (Ross, 1968). The decision of when to stop this interlocking of norms is left to the creator of norms. If a system requires it, the model (and formalisation) for enforcing and encouraging norms can be used recursively as necessary. There is nothing in the definition of the model itself to prevent this.

Both enforcement and reward norms acquire particular relevance in systems regulated by norms because the abilities to punish and reward must be restricted for use only by competent authorities (addressees of enforcement and reward norms). Otherwise, offenders might be punished twice or more if many agents take this as their responsibility. It could also be the case that selfish agents demand unjust punishments or that selfish offenders reject being punished. That is, conflicts of interest might emerge in a society if such responsibilities are given either to no one or to anyone. Only through enforcement and reward norms can agents become entitled to punish or reward other agents.

4.3 Legislation Norms

Norms are introduced into a society as a means to achieve social order. Some are intended to avoid conflicts between agents, others to allow the establish-

ment of commitments, and others still to unify the behaviour of agents as a means of social identification. However, neither all conflicts nor all commitments can be anticipated. Consequently, there must exist the possibility of creating new norms (to solve unexpected and recurrent conflicts among agents), modifying existing ones (to increase their effectiveness), or even abolishing those that become obsolete. As above, these capabilities must be restricted to avoid conflicts of interest. That is, norms stating when actions to legislate are permitted must exist in a normative multi-agent system (Jones and Sergot, 1996). Formally, we say that a norm is a *legislation* norm if actions to issue and to abolish norms are permitted by this norm in the current environment. These constraints are specified below.

$$
\begin{array}{|l}
legislate\_ : \mathbb{P}(Norm \times EnvState) \\
\hline
\forall\, n : Norm;\ env : EnvState \bullet \\
legislate\ (n, env) \Leftrightarrow \\
\quad (\exists\, issuingnorms, abolishnorms : Action \bullet \\
\qquad permitted\ (issuingnorms, n, env) \lor \\
\qquad permitted\ (abolishnorms, n, env)) \\
\end{array}
$$

4.4 Normative Multi-Agent Systems Model

A normative multi-agent system is formally represented in the $NormativeMAS$ schema. It comprises a set of normative agent members (i.e. agents able to reason about norms) and a set of general norms that govern the behaviour of these agents ($generalnorms$). Norms issued to allow the creation and abolition of norms ($legislationnorms$) are also included. There are also norms dedicated to enforcing other norms ($enforcenorms$) and norms directed to encouraging compliance with norms through rewards ($rewardnorms$). Legislation, enforcement and reward norms are better discussed in (López and Luck, 2004). The current state of the environment is represented by the variable $environment$. Constraints over these components are imposed as follows. Although it is possible that agents do not know all the norms in the system, it is always expected that they at least adopt some norms, represented by the first predicate. The second predicate makes explicit that addressees of norms must be members of the system. Thus, addressee agents of every norm must be included in the set of member agents because it does not make any sense to have norms addressed to nonexistent agents. The last three predicates respectively describe the structure of enforcement, reward and legislation norms. Notice that whereas every enforcement norm must have a norm to enforce, not every norm may have a corresponding enforcement norm, in which case no one in the society is legally entitled to punish an agent that does not fulfill such a norm.

$$
\begin{array}{|l}
NormativeMAS \\
\hline
members : \mathbb{P}\, NormativeAgent \\
generalnorms, legislationnorms : \mathbb{P}\, Norm \\
enforcenorms, rewardnorms : \mathbb{P}\, Norm \\
environment : EnvState \\
\hline
\forall\, ag : members \bullet \\
\quad ag.norms \cap generalnorms \neq \varnothing \\
\forall\, sn : generalnorms \bullet \\
\quad sn.addressees \subseteq members \\
\forall\, en : enforcenorms \bullet \\
\quad (\exists\, n : generalnorms \bullet enforces\ (en, n)) \\
\forall\, rn : rewardnorms \bullet \\
\quad (\exists\, n : generalnorms \bullet rewardnorm\ (rn, n)) \\
\forall\, ln : legislationnorms \bullet \\
\quad legislate\ (ln, environment) \\
\end{array}
$$

4.5 Normative Roles

Defining normative multi-agent systems in this way allows the identification of the *authorities* of the system as formalised in the $AuthoritiesNMAS$ schema. The set of agents that are entitled to create, modify, or abolish norms is called *legislators*. No other members of the society are endowed with this authority, and generally they are either elected or imposed by other agents. *Defender* agents are directly responsible for the application of punishments when norms are violated. That is, their main responsibility is to monitor compliance with norms in order to detect transgressions. Moreover, they can also warn agents by advertising the bad consequences of being rebellious. By contrast, *promoter* agents are those whose responsibilities include rewarding compliant addressees. These agents also monitor compliance with norms in order to determine when rewards must be given, and instead of *enforcing* compliance with norms, they simply *encourage* it.

$$
\begin{array}{|l}
AuthoritiesNMAS \\
\hline
NormativeMAS \\
legislators : \mathbb{P}\, NormativeAgent \\
defenders : \mathbb{P}\, NormativeAgent \\
promoters : \mathbb{P}\, NormativeAgent \\
\hline
\forall\, lg : legislators \bullet (\exists\, l : legislationnorms \bullet \\
\quad lg \in l.addressees) \\
\forall\, df : defenders \bullet (\exists\, e : enforcenorms \bullet \\
\quad df \in e.addressees) \\
\forall\, pm : promoters \bullet (\exists\, r : rewardnorms \bullet \\
\quad pm \in r.addressees) \\
\end{array}
$$

5 Autonomous Normative Reasoning

Whereas agents that always comply with norms are important for the design of societies in which total control is needed (Boman, 1999; Shoham and Tennenholtz, 1995), agents that can decide on the basis of their own goals and motivations whether to comply with them are important for the design of dynamic systems in which agents act on behalf of different users and, while satisfying their own goals, are able to join a society and cooperate with other agents. Autonomous norm reasoning is important to address those situations in which an agent's goals conflict with the norms that control its behaviour inside a society. Agents that deliberate about norms are also needed in systems in which unforseen events might occur, and in those situations in which agents are faced with conflicting norms, and they have to choose between them. It should be clear that violation of norms is, sometimes, justified. To describe *normative reasoning*, therefore, we have to explain not only what might motivate an agent to adopt, dismiss or complying with a norm, but also the way in which this decision affects its goals. In consequence we propose three different processes: one for agents to decide whether to adopt a norm (*the norm adoption process*), another to decide whether to comply with a norm (*the norm deliberation process*), and the other to update the goals, and therefore the intentions of agents accordingly (*the norm compliance process*). All these processes must take into account not only the goals and motivations of agents, but also the mechanisms of the society to avoid violation of norms such as rewards and punishments. Thus, agents consider the so called *social pressure of norms* before making any decision.

5.1 The Norm Adoption Process

The *norm adoption* process can be better defined as the process through which agents recognise their responsibilities towards other agents by internalising the norms that specify these responsibilities. Thus, agents adopt the norms of a society either once they have decided to join it or in the case a new norm is issued while they are still there. For autonomous agents to join and stay in a society the *social satisfaction condition* must hold (López et al., 2004). An agent considers this condition as satisfied if, although some of its goals become hindered by its *responsibilities*, its important goals can still be satisfied. Thus, we consider that the following conditions must be satisfied for agents to adopt a norm: the agent must recognise itself as an addressee of the norm; the norm must not already be adopted; the norm must have been issued by a recognised authority; and the agent must have reasons to stay in the society. Notice that to adopt a norm as an end, only the first three conditions are needed, whereas the last condition is an indicator that the decision to adopt a norm is made in an autonomous way. Due to space constraints, the $NormAdoption$ schema only formalises the first three conditions but details of the fourth condition can be found elsewhere (López et al., 2004).

$NormAdoption$

$$\Delta NormativeAgent$$
$$new? : Norm$$
$$issuer?, self : NormativeAgent$$
$$authorities : \mathbb{P}\, NormativeAgent$$
$$issuedby : \mathbb{P}(Norm \times NormativeAgent)$$

$$self \in new?.addressees$$
$$new? \notin norms$$
$$(new?,\ issuer?) \in issuedby \Leftrightarrow$$
$$\quad issuer? \in authorities$$
$$norms' = norms \cup \{new?\}$$

5.2 The Norm Deliberation Process

To comply with the norm, agents assess two things: the goals that might be hindered by satisfying the normative goals, and the goals that might benefit from the associated rewards. By contrast, to reject a norm, agents evaluate the damaging effects of punishments (i.e. the goals hindered due to the satisfaction of the goals associated with punishments.) Since the satisfaction of some of their goals might be prevented in both cases, agents use the *importance* of their goals to make these decisions. This, to deliberate about a norm, an agents pursues the following steps.

- A set of *active* norms is selected from the set of adopted norms (norm instances). Active norms are those that agents believe must be complied with in the current state, which is not an exception state (i.e. those norms for which the context matches the beliefs of the agent).
- The agent divides active norms into *non-conflicticting* and *conflicting* norms. An active norm is *non-conflicting* if its compliance does not cause any conflict with one of the agent's current goals. Thus, no goals of the addressee agent are hindered by satisfying the normative goals of the norm. By contrast, an active norm is *conflicting* if its fulfillment hinders any of the agent's goals.
- For each one of these sets of norms, the agent must decide which one to comply with. Details of different ways to select the norms to be

intended or rejected are given in (López et al., 2002). After norm deliberation, the set of intended norms consists of those conflicting and non-conflicting norms that are accepted to be complied with by the agent, and the set of rejected norms consists of all conflicting and non-conflicting norms that are rejected by the agent.

The state of an agent that has selected the norms it is keen to fulfill is formally represented in the $NormAgentState$ schema. This represents a normative agent with a variable representing the sets of $active$ norms at a particular point of time. The $conflicting$ predicate holds for a norm if and only if its normative goals conflict ($hinder$) with any of the agent's current goals. The next three predicates state that active norms are the subset of adopted norms that the agent believes must be complied with in the current state and that, the set of active norms has already been assessed and divided into norms to intend and norms to reject. The state of an agent is consistent in that its current goals do not conflict with the intended norms and, consequently, no normative goal must be in conflict with current goals. Moreover, since rewards benefit the achievement of some goals, so that agents do not have to work on their satisfaction because someone else does, these goals must not be part of the goals of an agent. The final predicate states that punishments must be accepted and, therefore, none of the goals of an agent must hinder them.

$NormAgentState$

$NormativeAgent$
$activenorms, conflicting\_ : \mathbb{P}\, Norm$

$\forall\, n : activenorms \bullet conflicting\ n \Leftrightarrow$
$\quad hinder(goals, n.ngoals) \neq \varnothing$
$activenorms \subseteq norms$
$\forall\, an : activenorms \bullet$
$\quad logcon\ (beliefs, an.context)$
$activenorms = intended \cup rejected$
$hinder(goals, normgoals\ intended) = \varnothing$
$benefit(goals, rewardgoals\ intended)$
$\quad \cap goals = \varnothing$
$hinder(goals, punishgoals\ rejected) = \varnothing$

For a norm to be intended, some constraints must be fulfilled. First, the agent must be an addressee of the norm. Then, the norm must be an adopted and currently active norm, and it must not be already intended. In addition, the agent must believe that it is not in an *exception* state and, therefore, it must comply with the norm. Formally, the process to accept a single norm as input ($new?$) to be complied with is specified in the $NormIntend$ schema. The first five predicates represent the constraints on the agent and the norm as described above. The sixth predicate represents the addition of the accepted norm to the set of intended norms and the final predicate represents the set of rejected norms remains the same.

$NormIntend$

$new? : Norm$
$\Delta NormAgentState$

$self \in new?.addressees$
$new? \in norms$
$new? \in activenorms$
$new? \notin intended$
$\neg\ logcon(beliefs, new?.exceptions)$
$intended' = intended \cup \{new?\}$
$rejected' = rejected$

The process to reject a norm ($NormReject$) can be defined similarly. Now, there are different ways to select the norms to be intended or rejected as explained in (López et al., 2002). Here, we describe what is called a *pressured* strategy where an agent fulfills a norm only in the case that one of its goals is threatened by punishments. That is, agents are *pressured* to obey norms through the application of punishments that might hinder some of their important goals. In this situation, the agent faces four different cases.

1. The norm is a non-conflicting norm and some goals are hindered by its punishments.
2. The norm is a non-conflicting norm and there are no goals hindered by its punishments.
3. The norm is a conflicting norm and the goals hindered by its normative goals are less important than the goals hindered by its punishments.
4. The norm is a conflicting norm and the goals hindered by its normative goals are more important than the goals hindered by its punishments.

The first case represents the situation in which, by complying with a norm, an agent does not put at risk any of its goals (because the norm is non-conflicting), but if the agent decides not to fulfill it, some of its goals could be unsatisfied due to punishments. Consequently, fulfilling a norm is the best decision for this kind of agent. To formalise this, we use the $NormIntend$ operation schema to accept complying with the norm, and we add two predicates to specify that this strategy is applied to non-conflicting norms whose punishments hinder some goals.

$PressuredNCComply$

$NormIntend$

$\neg\ conflicting\ new?$
$hinder(goals, new?.punishments) \neq \varnothing$

In the second case, by contrast, since punishments do not affect an agent's goals, it does not make any sense to comply with the norm, so it must be rejected. Formally, the $NormReject$ operation schema is used when the norm is non-conflicting (first predicate) and its associated punishments do not hinder any existing goals (second predicate).

$PressuredNCReject$

$NormReject$

$\neg$ *conflicting new?*
$hinder(goals, new?.punishments) = \varnothing$

According to our definition, a conflicting norm is a norm whose normative goals hinder an agent's goals. In this situation, agents comply with the norm at the expense of existing goals only if what they can lose through punishments is more important than what they can lose by complying with the norm. Formally, a conflicting norm is intended if the goals that could be hindered by punishments (hps) are more important than the set of existing goals hindered by normative goals ($hngs$). This is represented in the $PressuredCComply$ schema where the $importance$ function uses the motivations associated with the set of goals to find the importance of goals.

$PressuredCComply$

$NormIntend$

conflicting new?
let $hps == hinder(goals, new?.punishments)$ •
let $hngs == hinder(goals, new?.ngoals)$ •
$importance\ (motivations, hps) >$
$importance\ (motivations, hngs)$

However, if the goals hindered by normative goals are more important than the goals hindered by punishment, agents prefer to face such punishments for the sake of their important goals and, therefore, the norm is rejected. Formally, a conflicting norm is rejected by using the $NormReject$ operation schema if the goals hindered by its punishments (hps) are less important than the goals hindered by its normative goals ($hngs$).

$PressuredCReject$

$NormReject$

conflicting new?
let $hps == hinder(goals, new?.punishments)$ •
let $hngs == hinder(goals, new?.ngoals)$ •
$importance\ (motivations, hps) \leq$
$importance\ (motivations, hngs)$

All these cases are illustrated in Figure 4

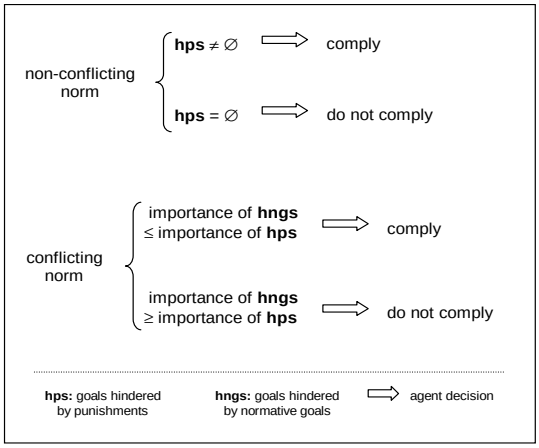

Figure 4: Pressured Norm Compliance

5.3 The Norm Compliance Process

Once agents take a decision about which norms to fulfill, a process of *norm compliance* must be started in order to update an agent's goals in accordance with the decisions it has made. An agent's goals are affected in different ways, depending on whether the norm is intended or rejected. The cases can be listed as follows.

- All normative goals of an intended norm must be added to the set of goals because the agent has decided to comply with it.
- Some goals are hindered by the normative goals of an intended norm. These goals can no longer be achieved because the agent prefers to comply with the norm and, consequently, this set of goals must be removed from the agent's goals.
- Some goals benefit from the rewards of an intended norm. Rewards contribute to the satisfaction of these goals without the agent having to make any extra effort. As a result, those goals that benefit from rewards must no longer be considered by the agent to be satisfied, and must be removed from the set of goals.
- Rejected norms only affect the set of goals hindered by the associated punishments. This set of goals must be removed; this is the way in which normative agents accept the consequences of their decisions.

To make the model simple, we assume that punishments are always applied, and rewards are always given, though the possibility exists that agents never become either punished or rewarded. In addition, note that the set of goals hindered by normative goals can be empty if the norm being considered is a non-conflicting norm, and goals hindered by punishments

or goals that benefit from rewards can be empty if a norm does not include any of them. After norm compliance, the goals are updated and, consequently, the intentions of agents might change. The process to comply with the norms an agent has decided to fulfill is specified in the $NormComply$ schema. Through this process, the set of goals is updated according to our discussion above.

$NormComply$

$\Delta NormAgentState$

---

let $ngs == \bigcup\{gs : \mathbb{P}\, Goal \mid (\exists\, n : intended \bullet gs = n.ngoals)\} \bullet$
let $hngs == \bigcup\{gs : \mathbb{P}\, Goal \mid (\exists\, n : intended \bullet gs = hinder\ (goals, n.ngoals))\} \bullet$
let $brs == \bigcup\{gs : \mathbb{P}\, Goal \mid (\exists\, n : intended \bullet gs = benefit(goals, n.rewards))\} \bullet$
let $hps == \bigcup\{gs : \mathbb{P}\, Goal \mid (\exists\, n : rejected \bullet gs = hinder\ (goals, n.punishments))\} \bullet$
$(\ goals' = (goals \cup ngs) \setminus (hngs \cup brs \cup hps))$

6 Conclusions

In this paper, we have presented a normative framework which, besides providing the means to computationally represent many normative concepts, can be used to give a better understanding of norms and normative agent behaviour. The framework explains not only the role that norms play in a society but also the elements that constitute a norm and that, in turn, can be used by agents when decisions concerning norms must be taken. In contrast to other proposals, our normative framework has been built upon the idea of *autonomy* of agents. That is, it is intended to be used by agents that reason about why norms must be adopted, and why an adopted norm must be complied with. Our framework consists of three main components: a canonical model of norms, a model of normative multi-agent systems and a model of normative autonomous agents.

The model of norms differs from others (Boman, 1999; Shoham and Tennenholtz, 1995; Tuomela, 1995) in the way in which patterns of behaviour are prescribed. To describe the pattern of behaviour prescribed by a norm, other models use actions, so that agents are told what exactly they must do. By contrast, we use normative goals, which is an idea more compatible with autonomous agents whose behaviour is driven by goals. Agents can choose the way to satisfy the normative goals, instead of being told exactly how it must be done. Our work also emphasises that all norms can be represented by using similar components, and that they are analysed by agents in similar ways. However, what makes one norm different from another is the way in which norms are created, how long they are valid, and the reasons agents have to adopt them. These factors enable norms to be divided into categories such as obligations and prohibitions, social commitments and social codes.

A collateral result of our work is the proposed model for interlocking norms. These relations between norms have already been mentioned in several papers, especially from philosophical and legal perspectives (Ross, 1968), but no ways to model them have been provided. Dignum's concept of authorisations (Dignum, 1999) attempts to describe norms activated when others are not fulfilled; however, his idea and models are incomplete. We claim that this form of representing connections between norms can be used not only to represent enforcement and reward norms, but also to represent things as complex as contracts and deals among agents.

In contrast to current models of systems regulated by norms (Balzer and Tuomela, 2001; Dignum and Dignum, 2001; Esteva et al., 2001; Shoham and Tennenholtz, 1995) in which no distinction among norms is made, our work emphasises that besides the general norms of the system, at least three kinds of norms are needed, namely norms to legislate, to punish, and to reward other agents. By making this differentiation, agents are able to determine when an issued norm is valid, when an entitled agent can apply a punishment, and who is responsible for giving rewards. In addition, order is imposed on agents responsible for the normative behaviour of other agents, because their authority is defined by the norms that entitle them to exert social pressure. Roles for *legislators*, *defenders*, and *promoters* of norms become easily identified as a consequence of the different kinds of norms considered. Thus, in this framework, the authority of agents is always supported and constrained by norms.

Acknowledgements

The first author acknowledges funding from the Faculty Enhancement Program (PROMEP) of the Mexican Ministry of Public Education (SEP), project No PROMEP/103.5/04/767.

References

R. Axelrod. An evolutionary approach to norms. *The American Political Science Review*, 80(4):1095–1111, 1986.

W. Balzer and R. Tuomela. Social institutions, norms and practices. In C. Dellarocas and R Conte, editors, *Social Order in Multi-Agent Systems*, pages 161–180. Kluwer Academic, 2001.

M. Barbuceanu, T. Gray, and S. Mankovski. The role of obligations in multiagent coordination. *Applied Artificial Intelligence*, 13(1/2):11–38, 1999.

G. Boella and L. Lesmo. Deliberative normative agents. In C. Dellarocas and R Conte, editors, *Social Order in Multi-Agent Systems*, pages 85–110. Kluwer Academic, 2001.

M. Boman. Norms in artificial decision making. *Artificial Intelligence and Law*, 7(1):17–35, 1999.

C. Castelfranchi, F. Dignum, C. Jonker, and J. Treur. Deliberative normative agents: Principles and architecture. In N. Jennings and Y. Lesperance, editors, *Intelligent Agents VI*, LNAI 1757, pages 206–220. Springer, 2000.

R. Conte and C. Castelfranchi. *Cognitive and Social Action*. UCL Press, 1995.

R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In J. Müller, M. Singh, and A. Rao, editors, *Intelligent Agents V*, LNAI 1555, pages 319–333. Springer, 1999a.

R. Conte and Ch. Dellarocas. Social order in info societies: An old challenge for innovation. In C. Dellarocas and R Conte, editors, *Social Order in Multi-Agent Systems*, pages 1–15. Kluwer Academic, 2001.

R. Conte, R. Falcone, and G. Sartor. Agents and norms: How to fill the gap? *Artificial Intelligence and Law*, 7(1):1–15, 1999.

C. Dellarocas and M. Klein. Contractual agent societies: Negotiated shared context and social control in open multi-agent systems. In C. Dellarocas and R Conte, editors, *Social Order in Multi-Agent Systems*, pages 113–133. Kluwer Academic, 2001.

F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1):69–79, 1999.

F. Dignum, D. Morley, E. Sonenberg, and L. Cavendon. Towards socially sophisticated BDI agents. In Edmund H. Durfee, editor, *The Fourth International Conference on Multi-Agent Systems*, pages 111–118. IEEE Computer Society, 2000.

V. Dignum and F. Dignum. Modelling agent societies: Coordination frameworks and institutions. In P. Brazdil and A. Jorge, editors, *Progress in Artificial Intelligence Knowledge Extraction, Multi-agent Systems, Logic Programming, and Constraint Solving*, LNAI 2258, pages 191–204. Springer-Verlag, 2001.

M. d'Inverno and M. Luck. *Understanding Agent Systems*. Springer-Verlag, second edition, 2003.

M. Esteva, J. Padget, and C. Sierra. Formalizing a language for institutions and norms. In J. Meyer and M. Tambe, editors, *Intelligent Agents VIII*, LNAI 2333, pages 348–366. Springer, 2001.

A. Jones and M. Sergot. A formal characterisation of institutionalised power. *Logic Journal of the IGPL*, 4(3):429–445, 1996.

F. López y López and M. Luck. Modelling norms for autonomous agents. In E. Chávez, J. Favela, M. Mejía, and A. Oliart, editors, *The Fourth Mexican Conference on Computer Science*, pages 238–245. IEEE Computer Society, 2003.

F. López y López and M. Luck. A model of normative multi-agent systems and dynamic relationships. In G. Lindemann, D. Moldt, and M. Paolucci, editors, *Regulated Agent-Based Social Systems*, LNAI 2934, pages 259–280. Springer-Verlag, 2004.

F. López y López, M. Luck, and M. d'Inverno. Constraining autonomy through norms. In C. Castelfranchi and W.L. Johnson, editors, *The First International Joint Conference on Autonomous Agents and Multi Agent Systems AAMAS'02*, pages 674–681. ACM Press, 2002.

F. López y López, M. Luck, and M. d'Inverno. Normative agent reasoning in dynamic societies. In N. Jennings, C. Sierra, L. Sonenberg, and L Tambe, editors, *The Third International Joint Conference on Autonomous Agents and Multi Agent Systems AAMAS'04*, pages 730–737. ACM Press, 2004.

M. Luck, P. McBurney, and C. Preist. *Agent Technology: Enabling Next Generation Computing (A Roadmap for Agent Based Computing)*. AgentLink, 2003.

T. Norman, C. Sierra, and N. Jennings. Rights and commitments in multi-agent agreements. In Yves Demazeau, editor, *The Third International Conference on Multi-Agent Systems (ICMAS-98)*, pages 222–229. IEEE Computer Society Press, 1998.

A. Ross. *Directives and Norms*. Routledge and Kegan Paul Ltd., 1968.

Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: Off-line design. *Artificial Intelligence*, 73(1-2):231–252, 1995.

M. Singh. An ontology for commitments in multiagent systems: Toward a unification of normative concepts. *Artificial Intelligence and Law*, 7(1):97–113, 1999.

J. M. Spivey. *The Z Notation: A Reference Manual*. Prentice-Hall, 1992.

R Tuomela. *The Importance of Us: A Philosophical Study of Basic Social Norms*. Stanford University Press, 1995.