Constraining Autonomy through Norms

Fabiola López y López
Electronics and Computer Science
Southampton University
Southampton, UK
¤yl00r@ecs.soton.ac.uk

Michael Luck
Electronics and Computer Science
Southampton University
Southampton, UK
mml@ecs.soton.ac.uk

Mark d'Inverno
Cavendish School of Computer Science
Westminster University
London, W1M 8JS, UK
dinverm@westminster.ac.uk

ABSTRACT

Despite many efforts to understand why and how norms can be incorporated into agents and multi-agent systems, there are still several gaps that must be filled. This paper focuses on one of the most important processes concerned with norms, namely that of *norm compliance*. However, instead of taking a static view of norms in which norms are straighforwardly complied with, we adopt a more dynamic view in which an agent's motivations, and therefore its autonomy, play an important role. We analyse the motivations that an agent might have to comply with norms, and then formally propose a set of strategies for use by agents in norm-based systems. Finally, through some simulation experiments, the effects of autonomous norm compliance in both individual agents and societies are analysed.

Categories and Subject Descriptors

I.2.11 [**ARTIFICIAL INTELLIGENCE**]: Distributed Artificial Intelligence—*Intelligent Agents*

General Terms

Theory, Experimentation

Keywords

Social order, control, norms

1. INTRODUCTION

It has been argued by many [3, 4, 7, 18] that agents working in a common society need to be constrained in order to avoid and solve conflicts, make agreements, reduce complexity, and in general to achieve a desirable social order. This is the role of norms, which represent what ought to be done by a set of agents, and whose fulfillment can be generally seen as a public good when their benefits can be enjoyed by the overall society, organisation or group [2]. Indeed, norms represent the means to achieve the goals of a society, and therefore their study becomes interesting. Research on norms and agents has ranged from fundamental work on the importance of norms in agent behaviour [7, 21] to proposing internal representations of norms [5, 22], considering their emergence in groups of agents [23], and proposing logics for their formalisation [19, 24]. Despite such efforts to understand how and why norms can be incorporated into agents and multi-agent systems, there is still much work to do.

The easiest way to represent and reason about norms is by seeing them as built-in constraints where all the restrictions and obligations of agents are obeyed absolutely without deliberation. In this view, the effort is left to the system designer to ensure that all agents respond in the required way and, consequently, that the overall system behaves coherently. However, this may result in inflexible systems that must be changed off-line when either the agents or the environment change. By contrast, if a dynamic view of norms is taken, the flexibility of the overall system could be guaranteed [25]. Towards this end, agents must be endowed with abilities, first, to adopt new norms and then to comply with them. By introducing agents able to *adopt norms*, we allow the representation of multi-agent systems composed of heterogeneous agents, independently designed, which can dynamically belong to different societies (or multiple societies) with the ability to adopt different roles [12]. This is an useful property for working in virtual organizations, coalitions and human society simulations. Moreover, if this process is autonomous, agents may also have the possibility of selecting the society to which they want to belong based on their own motivations and preferences.

Turning to the process of *norm compliance*, agents can be represented as either entities that always comply with their norms, or entities that autonomously choose whether to do so or not. Both possibilities may cause conflicts between a society and the individuals within it. On the one hand, if norm compliance is assumed, social goals (achieved through norm obedience) are guaranteed. However, personal goals may be frustrated by obeying all the imposed norms because agents may lose opportunities that new situations offer to their individual interests. On the other hand, if the decision of whether to comply with a norm is left to the agent, although personal interest may be satisfied, the system becomes unpredictable when not all norms are obeyed, and consequently the society performance may be degraded. In this situation, enforcement mechanisms can be introduced as a means of persuading agents to obey the norms. That is, agents might comply with norms in order to either avoid a punishment or obtain a reward. As a result, we argue that a representation of agents able to deal with norm adoption

AAMAS'02, July 15-19, 2002, Bologna, Italy.

and compliance, as well as with the sanctions and rewards associated with them, is needed. Moreover, both adoption and fulfillment of norms are important decision processes where agent autonomy plays a significant role.

Now, many questions are posed: what does an agent take into account to decide whether to fulfill a norm; are some kinds of agents more appropriate for being part of a society; can a society frustrate individual agent aspirations; which kinds of agents are more affected by the fulfillment of norms; and so on. In consequence, the main purposes of this paper are, firstly, to describe how the process for autonomous norm compliance can be achieved, and secondly, to analyse its impact on both the society and the individuals within it. To this end, we start by defining norms and normative agents in the next section. After that, autonomous norm compliance, enforcement mechanisms and the agent motivations for fulfilling norms are analysed. Then, the formal model for both the norm compliance process and the strategies for norm decision-making are described. In providing answers to some of the questions above, experiments based on these processes, and their results, are described. Finally, both conclusions and future work are provided.

2. NORMS AND NORMATIVE AGENTS

As a means of building up a formal model of a normative agent without being repetitive, we adopt the SMART *agent framework* described in [10]. In addition, in what follows, we use the Z specification language to construct such a formal model. Z is based on set-theory and first order logic, with details available in [20]. For brevity, however, we will not elaborate the use of Z further.

2.1 Agents

In the SMART *agent framework*, an *attribute* represents a perceivable feature of the agent's environment, *goals* are defined as a non-empty set of attributes that describe states of affairs in the world, *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, further details are not needed, so we simply consider them as given sets.

$$[Attribute, Goal, Motivation, Action]$$

In addition, an entity is described by a non-empty set of attributes representing its permanent features, a set of goals that it wants to bring about, a set of capabilities that it is able to perform, and a set of motivations representing its preferences. Moreover, *agents* are entities whose set of goals is not empty, and *autonomous agents* are agents with non-empty sets of motivations. By omitting irrelevant details, we formalise them as follows.

Agent

$capabilities : \mathbb{P}\, Action$
$goals : \mathbb{P}\, Goal$
$motivations : \mathbb{P}\, Motivation$
$beliefs : \mathbb{P}\, Attribute$

$goals \neq \varnothing$

$$AutonomousAgent \mathrel{\widehat=} [Agent \mid motivations \neq \varnothing]$$

2.2 Norms

It can be said that *norms* are mechanisms that a society has in order to influence the behaviour of agents within it. Norms can be created from different sources, varying from built-in norms to simple agreements between agents, or more complex legal systems. They may persist during different periods of time, for example until an agent dies, as long as an agent remains in the society for which the norms were issued, or just for a short period of time until a normative goal becomes satisfied. There are different aspects that can be used for characterizing them. First, norms are always prescribed to be complied with for a set of *addressee* agents in order to *benefit* another set of agents (possibly empty). They specify something that ought to be done, and consequently they include *normative goals* that must be satisfied by addressees. Sometimes, these normative goals must be directly intended, whereas other times their role is to inhibit specific goals (as in the case of prohibitions). Second, norms are not always applicable, and their activation depends on the *context* in which agents are situated. Moreover, there may be *exception* states where agents are not obliged to comply with the norm. Finally, in some cases, norms suggest the existence of a set of *sanctions* or *punishments* to be imposed when agents do not satisfy the normative goal, and a set of *rewards* to be received when agents do. Thus, the general structure of a norm can be formalised as follows. (Note that we specify normative goals as a set, to allow for the possibility of multiple goals in a norm, though we recognise that this will typically be a singleton set.)

Norm

$addressees, beneficiaries : \mathbb{P}\, Agent$
$normativegoals, rewards, punishments : \mathbb{P}\, Goal$
$context, exceptions : \mathbb{P}\, Attribute$

$addressees \neq \varnothing$
$context \neq \varnothing$

Norms can be divided, without eliminating the possibility of having further categories, into four types: *obligations*, *prohibitions*, *social commitments* and *social codes*. Roughly, we can say that *obligations* and *prohibitions* are norms adopted once an agent becomes a member of a society, *social commitments* are norms derived from agreements or negotiations between two or more agents, and *social codes* are norms motivated to be followed by feelings such as love, pity, friendship, or social conformity. It is not the purpose of this paper to discuss the different categories of norms; consequently, in the remainder of this document we will use the term *norm* as an umbrella term to cover every type of norm. However, we state that all of them share the same structure.

2.3 Normative Agents

In general, a normative agent is an autonomous agent whose behaviour is shaped by the obligations it must comply with, prohibitions that limit the kind of goals that it can pursue, social commitments that have been created during its social life and social codes which may not carry punishments, but whose fulfillment could represent social satisfaction for the agent.

NormativeAgent

$AutonomousAgent$
$norms : \mathbb{P}\, Norm$

3. AUTONOMOUS NORM COMPLIANCE

It is important to mention that *adoption of* and *compliance with* norms are two different, but related, processes. The first involves the agent's acknowledgment of three facts: it is part of the society, it is an addressee of the norms, and the issuer of the norm is someone entitled to do it. By contrast, the compliance with norms involves an agent's commitment to obey the norm and therefore to achieve the associated normative goals. During the norm adoption process, norms are recognised as duties by the agent. It knows the norm, and the majority of the times it is willing to obey it. However, at run time the situation of an agent may change, making it difficult to maintain its compromise of obeying the norm, especially if that norm is causing conflict with its individual goals. Therefore, before complying with a norm, an agent must evaluate whether its fulfillment will satisfy its personal current motivations and preferences. In other words, an autonomous agent must not only decide which goals to pursue, how these goals can be achieved and which external goals can be adopted [15, 16], it also must decide which norms to fulfill, based on its motivations. Sometimes norms are obeyed as an end just because agents have intrinsic motivations to be social. Other times, agents only obey norms if a punishment is applied for not doing so, if they are rewarded, and others still are guided by their internal motivation to be trusted. However, norms are sometimes violated, and to understand why, we must also analyse the motivations agents have to do so.

3.1 Enforcement Mechanisms

Some *enforcement mechanisms* are needed as a means of ensuring that personal interests do not overcome social rules. Usually enforcement mechanisms are associated with punishments and rewards so that agents are obliged to obey norms because of either the fear of being punished or the desire to gain something. However, as some sociologists point out [11], punishments and rewards will only affect an agent's decision to comply with norms if they either hinder or benefit one of the agent's goals. That is, punishments cannot be taken into account if none of the agent's interests (translated as individual goals) is going to be hindered. For example, the norm of wearing fashionable clothes may have an associated punishment of not being socially accepted. However, this applies just to a specific group of agents, and there may be others less interested in being accepted, who therefore consider the fulfillment of that norm as unworthy. Rewards are similarly a means to motivate agents only if one of the agent's goals receives benefits from such fulfillment. Therefore, we can say that punishments and rewards do not have any effect on an agent's decision if they are not associated with some of the agent's individual goals.

Now, since punishments and rewards are defined as goals, and in order to know their effects on an agent's overarching goals, we need to understand when a goal can either hinder or benefit another one. In general, a goal can hinder another one when they are in conflict. Sometimes such a conflict is easy to observe because the state of one goal is simply the negation of the other, such as being outside a room and inside it at the same time. However, conflicting situations in general are more difficult to observe. For example, cleaning up a room and watching a favourite TV programme can be in conflict if they are intended at the same time and in different places. Goals receiving benefits are similar in that the easiest way to observe situations where a goal benefits from another is when both goals represent the same state but achieved by different agents. Representing goals in conflict and beneficial goals are beyond the scope of this paper; therefore, we take them as given direct associations between two goals without giving further details.

Two functions to get all the goals from a set of goals that can be either hindered by, or benefited from, a second set of goals can be formally defined as below.

$$\begin{array}{|l} related : (\mathbb{P}\, Goal \times \mathbb{P}\, Goal \times (Goal \leftrightarrow Goal)) \\ \qquad \twoheadrightarrow \mathbb{P}\, Goal \\ \hline \forall\, gs_1, gs_2 : \mathbb{P}\, Goal;\ rtd : Goal \leftrightarrow Goal \bullet \\ \quad related\ (gs_1, gs_2, rtd) = \mathrm{ran}(gs_2 \lhd (rtd \rhd gs_1)) \end{array}$$

Both the functions *hindered* and *benefited* have this specification.

$$hindered == related; \qquad benefited == related$$

3.2 Motivations for Norm Compliance

In general, norms are broken when their fulfillment may hinder personal goals that agents consider as worthy for their personal interest, or when agents have internal motivations to reject external orders. Whatever the causes to violate a norm, both society and individuals may be affected. On the one hand, societies issue norms to be obeyed as mechanisms to achieve social goals, and it is expected that all society members comply with them. On the other hand, agents have their own goals which may be frustrated in order to comply with their duties. For example, in the case of the obligation of paying taxes, the society as a whole expects the norm to be fulfilled because it is a means to achieve social welfare, but such an obligation may frustrate the personal goals of taking abroad holidays or buying something. In this case, the decision concerns only the agent which, based on its motivations and current situation, must decide what is more important for it. Some careless agents may take this decision just by considering both the normative and their personal goals, but others may also take into consideration the consequences of being either punished or rewarded. For example, if an agent decides do not pay its taxes and continues with its goal towards some enjoyable holidays, it must accept the consequences of being punished. Conversely, if agents are cautious they must consider both the possibility of being punished and how much the punishment may affect their other personal goals. In general, agents comply with norms in several ways, which are listed below.

- When an agent is strongly motivated by its social concerns, and its social goals are more important than personal goals, all norms are fulfilled, even though some of its goals are hindered. In this situation, we say that an agent is being *social*.
- Sometimes the fulfillment of a norm is considered as last resort in order to avoid some personal goals becoming prevented by sanctions. In other words, agents are *pressured* to obey norms by applying punishments that might hinder some of their important goals.
- There are also *opportune* situations where the fulfillment of a norm may contribute to the achievement of one of an agent's goals. That is, compliance with norms is ensured through the benefits obtained from the rewards.

- The *fear* to be punished can also be considered here. However, contrary to the pressured form of norm compliance, fearful agents comply with norms even if the sanction does not affect any of their goals.
- *Greed* is also a motive for norm compliance. Greedy agents obey norms only if they receive something in exchange, even though none of their goals benefit from the reward.

Combinations of these strategies are also possible. For example, an agent can be pressured and opportunistic and therefore selfish because it only fulfills a norm when one of its interests becomes affected (hindered/benefited). Finally, we also may represent the situation when agents reject social norms as follows.

- In *rebellious* behaviour, agents may refuse to obey external orders even though by neglecting norms some of their goals could be hindered. Indeed, this is a kind of anti-social behaviour.

4. THE FORMAL MODEL

We adopt the vision of *motivated agency* proposed by Luck and d'Inverno [14, 17] in which an agent's preferences are expressed through motivations. In addition, we require that all goals and norms that agents have are motivated, meaning that agents have reasons to pursue goals as well as reasons for adopting norms. By using the motivations associated with a set of goals, their importance, for the agent being considered, can be found. This is expressed by the following function.

$$importance : \mathbb{P}\, Motivation \rightarrow \mathbb{P}\, Goal \rightarrow \mathbb{N}$$

In this way, the fact that not all goals have the same importance (or motivation) for agents is represented. At run time, these values are used to decide which goals should be achieved first. In this context, we can use the same values for deciding which goals an agent prefers to hold, because norm compliance may mean that some personal goals do not become satisfied, especially if there is a conflict with normative goals. In addition, to make our model simple, we assume that all punishments and rewards are applied by someone else. Therefore, the possibility of cheating the bearer of the norm [1] will not be considered here, and the application of punishments and rewards is taken for granted.

4.1 Norm Compliance Processes

Agents in this model are *normative agents* with the ability to choose the set of norms they want to comply with, to satisfy those norms, and to accept the consequences of not complying with norms. The process of norm compliance involves two sets of norms: the set of *active* norms (*activenorms*) which represents all currently active norms considered by an agent, and the set of *intended* norms (*intendednorms*) which represents those norms that the agent has decided to comply with. This latter is a subset of active norms, and it is different for each agent since it depends on the particular norm compliance strategy adopted as a result of an agent's motivations.

Now, we assume that the state of an agent is consistent in that its current goals do not conflict with the intended norms. An agent must thus know which goals are in conflict and which goals can benefit from other goals, represented by the *hinders* and *benefits* variables defined in Section 3.1. In the schema *NormativeAgentState*, we define an agent with no conflicting goals and norms, and include some new components. First, *rejectednorms* represents those norms that an agent does not intend. Second, *conflicting* is a predicate that holds for a norm if and only if the goal of the norm conflicts with any of the agent's current goals. This will be useful later when specifying the different types of agent. In addition, a set of useful functions to extract normative goals, punishments and rewards respectively are defined as follows.

$$normgoals, punishgoals, rewardgoals : \mathbb{P}\, Norm \rightarrow \mathbb{P}\, Goal$$

$$\begin{aligned}
&\forall\, ns : \mathbb{P}\, Norm \bullet \\
&\quad normgoals\ ns = \bigcup \{n : ns \bullet n.normativegoals\} \land \\
&\quad punishgoals\ ns = \bigcup \{n : ns \bullet n.punishments\} \land \\
&\quad rewardgoals\ ns = \bigcup \{n : ns \bullet n.rewards\}
\end{aligned}$$

NormativeAgentState

$$\begin{aligned}
&NormativeAgent \\
&activenorms : \mathbb{P}\, Norm \\
&intendednorms : \mathbb{P}\, Norm \\
&rejectednorms : \mathbb{P}\, Norm \\
&hinders : Goal \leftrightarrow Goal \\
&benefits : Goal \leftrightarrow Goal \\
&conflicting\ _ : \mathbb{P}\, Norm \\
\hline
&activenorms = intendednorms \cup rejectednorms \\
&hindered(goals, punishgoals\ rejectednorms, \\
&\qquad hinders) = \varnothing \\
&hindered(goals, normgoals\ intendednorms, \\
&\qquad hinders) = \varnothing \\
&benefited(goals, rewardgoals\ intendednorms, \\
&\qquad benefits) \cap goals = \varnothing \\
&\forall\, n : activenorms \bullet conflicting\ n \Leftrightarrow \\
&\quad hindered(goals, n.normativegoals, \\
&\qquad hinders) \neq \varnothing
\end{aligned}$$

The *norm compliance* process for a single norm as input (*new*?), which is shown in Schema *NormComply*, can be described as follows. The new norm is added to the set of intended norms, while the set of rejected norms remains the same. The set of current goals must be updated, so that the set of normative goals corresponding to the new accepted norm is added to them. As a result of this, any existing goals that conflict with these normative goals are hindered, and must be removed. Moreover, goals that benefit from the rewards associated with the new norm are also removed as a result of being achieved by other means. (Actually, this is a simplification, and the hindered goals may in fact simply be *suspended*, while the goals that benefit may be achieved after the norm has been fulfilled. For now, such issues complicate our presentation, and we omit a consideration of them in this paper.) In addition, the schema for not complying with an active norm (*NormNonComply*) is defined similarly, but in this case, punishments are incurred (gs_1), and therefore accepted by removing the goals they hinder. Notice that the value of *new*? is generated internally when considering a norm, though we specify it as an external input to the operations for now, and we remain neutral on which operation the agent chooses. The schema below introduces a new predicate, *logicalconsequence*, which is true when the second argument is a *logical consequence* of the first.

$logicalconsequence\_ : \mathbb{P}(\mathbb{P}\ Attribute \times \mathbb{P}\ Attribute)$

NormComply

$new? : Norm$
$\Delta NormativeAgentState$

$new? \in norms$
$new? \in activenorms$
$new? \notin intendednorms$
$logicalconsequence(beliefs, new?.context)$
$\neg\ logicalconsequence(beliefs, new?.exceptions)$
$intendednorms' = intendednorms \cup \{new?\}$
$rejectednorms' = rejectednorms$
let $gs_1 == new?.normativegoals \bullet$
let $gs_2 == hindered\ (goals, new?.normativegoals, hinders) \bullet$
let $gs_3 == benefited(goals, new?.rewards, benefits) \bullet$
$goals' = (goals \cup gs_1) \setminus (gs_2 \cup gs_3)$

NormNonComply

$new? : Norm$
$\Delta NormativeAgentState$

$new? \in norms$
$new? \in activenorms$
$new? \notin intendednorms$
$logicalconsequence(beliefs, new?.context)$
$\neg\ logicalconsequence(beliefs, new?.exceptions)$
$intendednorms' = intendednorms$
$rejectednorms' = rejectednorms \cup \{new?\}$
let $gs_1 == hindered\ (goals, new?.punishments, hinders) \bullet goals' = goals \setminus gs_1$

4.2 Strategies for Norm Compliance

Although norm compliance *process* is similar in all agents, different strategies can be used to find the set of *intended* norms, depending on what is considered to be important by an agent. To find it, the set of *active norms* is divided and analysed in two disjoint sets of norms: the first one including all norms which compliance does not cause any conflict with one of the agent's current goals, and the second including all active norms which fulfillment may hinder any of them. In that way, we are allowing the possibility of applying different strategies for each set of norms. In general, there are four sets of goals that must be observed to decide which norms to fulfill: the normative goals derived from the norm that is being considered, the agent's goals that could be hindered by this set of normative goals, the agent's goals that might be hindered by applying punishments, and the agent's goals that might be benefited by rewards. The set of goals hindered by normative goals could be empty if the norm being considered is a non-conflicting norm. In the same way goals hindered/benefited by punishments/rewards can also be empty if a norm does not include them. Now, according to the motivations for norm compliance described in Section 3 the possible strategies are listed below.

Social A social strategy can be adopted when social goals are more important than personal goals. Consequently, social agents will never be punished and will receive the maximum social benefits provided by rewards. However, this can result in the loss of a considerable number of existing goals if the normative goals conflict with them. Formally, we can simply state no norms are rejected.

SocialAgent

NormComply

$rejectednorms = \varnothing$

Pressured Agents adopting this strategy consider the effects of punishments on their existing goals and act accordingly. We can enumerate four distinct cases as follows. First, a non-conflicting norm is complied with only if the punishment hinders an existing goal.

PressuredAgentNNCComply

NormComply

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

If the punishment of a non-conflicting norm does not hinder any existing goals, the norm is rejected.

PressuredAgentNNCReject

NormNonComply

$\neg\ conflicting\ new?$
$hindered(goals, new?.punishments, hinders) = \varnothing$

In the case of conflicting norms, an agent will comply with the norm at the expense of existing goals only if the goals hindered by the punishments are more important than the set of existing goals hindered by normative goals.

PressuredAgentNCComply

NormComply

$conflicting\ new?$
let $gs_1 == hindered(goals, new?.punishments, hinders) \bullet$
let $gs_2 == hindered(goals, new?.normativegoals, hinders) \bullet$
$importance\ motivations\ gs_1 > importance\ motivations\ gs_2$

Otherwise, the agent rejects with the norm.

PressuredAgentNCReject

NormNonComply

$conflicting\ new?$
let $gs_1 == hindered(goals, new?.punishments, hinders) \bullet$
let $gs_2 == hindered(goals, new?.normativegoals, hinders) \bullet$
$importance\ motivations\ gs_1 \leq importance\ motivations\ gs_2$

Opportunistic In this strategy, agents consider the effects of rewards on their existing goals. Formally, this is similar to the *pressured* case. Non-conflicting norms are fulfilled only if their rewards benefit some goals.

OpportunisticAgentNNCComply
NormComply

$\neg$ *conflicting new*?
$benefited(goals, new?.rewards, benefits) \neq \varnothing$

OpportunisticAgentNNCReject
NormNonComply

$\neg$ *conflicting new*?
$benefited(goals, new?.rewards, benefits) = \varnothing$

In the case of conflicting norms, motivations again determine how to act. That is, conflicting norms are complied with only when their associated rewards benefit goals which are more important than those goals hindered by normative goals.

OpportunisticAgentNCComply
NormComply

conflicting new?
let $gs_1 == benefited(goals,$
$new?.rewards, benefits) \bullet$
let $gs_2 == hindered(goals,$
$new?.normativegoals, hinders) \bullet$
$importance\ motivations\ gs_1 >$
$importance\ motivations\ gs_2$

OpportunisticAgentNCReject
NormNonComply

conflicting new?
let $gs_1 == benefited(goals,$
$new?.rewards, benefits) \bullet$
let $gs_2 == hindered(goals,$
$new?.normativegoals, hinders) \bullet$
$importance\ motivations\ gs_1 \leq$
$importance\ motivations\ gs_2$

Fearful A fearful strategy means that an agent decides to comply with a norm only if it includes a punishment. No further deliberation is made here.

FearfulAgent
NormComply

$\forall\, n : rejectednorms \bullet n.punishments = \varnothing$
$\forall\, n : intendednorms \bullet n.punishments \neq \varnothing$

Greedy In a somewhat symmetric manner to fearful agents, the greedy strategy is adopted just for the pleasure of getting something, even it does not contribute to the agent's existing goals.

GreedyAgent
NormComply

$\forall\, n : rejectednorms \bullet n.rewards = \varnothing$
$\forall\, n : intendednorms \bullet n.rewards \neq \varnothing$

Rebellious Rebellious agents simply reject all norms.

RebelliousAgent
NormNonComply

$intendednorms = \varnothing$

Given that more than one strategy can be applied to each set of active norms, complex processes of norm compliance can be represented. For example, *selfish* norm compliance requires both pressured and opportunistic strategies and can be defined as follows. Further combinations of these are also possible, but we shall not explore them further here.

$Selfish == (PressuredAgentNNCComply \lor$
$PressuredAgentNNCReject \lor$
$PressuredAgentNCComply \lor$
$PressuredAgentNCReject)^{\S}_{9}$
$(OpportunisticAgentNNCComply \lor$
$OpportunisticAgentNNCReject \lor$
$OpportunisticAgentNCComply \lor$
$OpportunisticAgentNCReject)$

5. AGENT AND SOCIETY PERFORMANCE

In order to understand the impacts of norms on both individual agents and societies, we have developed a workbench in which the behaviour of normative agents can be tested and observed. Focusing exclusively on norm compliance effects, we assume a set of agents having similar capabilities, and being controlled by the same set of norms. Despite such similarities, each agent differs in the strategies for norm compliance that it chooses. In this workbench, different data can be monitored over time. For example, we can record the number of norms that become active during a specific period (*active norms*), the number of these norms that an agent complies with (*intended norms*), the number of an agent's current goals generated through motivations, those not hindered by norms, and those that benefit from rewards. However, to provide more useful information, we define the following quantitative measures. First, the *social contribution* of an agent can be defined as the number of times the agent complies with its responsibilities (expressed through norms) in proportion to the total number of active norms. Second, an agent's *individual satisfaction* is the number of personal satisfied goals as a proportion of its total number of goals generated over the same period. Thus, individual satisfaction represents those personal goals not hindered by either normative goals or punishments. Though these *have been* defined formally, space constraints prohibit the inclusion of the formal definitions here.

5.1 The Experimental Model

Our initial experimentation includes six kinds of agents following the strategies shown in Table 1, where *NCN* and *CN* represent *non-conflicting* and *conflicting* norms respectively. We choose these examples because they represent the most common strategies followed by agents when they face a norm compliance decision. The first two represent extreme behaviours, that is agents always obeying norms or agents always refusing them. The third case represents selfish agents who comply with norms only if such a decision provides them with benefits (through rewards), or because the punishment for not complying with them might cause negative effects on their goals. Social selfish agents obey norms only if they do not conflict with their own goals, otherwise they apply a selfish strategy. The last two kinds of agents are also variations of selfish agents, where agents apply either pressured or opportunistic strategies for non-conflicting norms, and a selfish strategy in other cases. In fact, several other combinations of strategies and norms have

been both modelled and tested, but space constraints restrict us to presenting a subset. Now, our interest here is to observe how both the *social contribution* and the *individual satisfaction* of each agent change according to both the strategy for norm compliance it chooses, and the increase in the number of conflicts between the norms it has to comply with and its personal goals.

Agent	Strategies for NCN	Strategies for CN
Social	Social	Social
Rebellious	Rebellious	Rebellious
Selfish	Pressured & Opportunistic	Pressured & Opportunistic
SocialSelf	Social	Opportunistic & Pressured
PressSelfish	Pressured	Opportunistic & Pressured
OppSelfish	Opportunistic	Opportunistic & Pressured

Table 1: Normative Agents Examples

Before a test is made, some variables are fixed as follows. First, a base of goals to represent all the goals that an agent might have is randomly created. Second, a random motivation value is associated to each goal in this set to represent the importance of each goal. In addition, each goal is used as a normative goal to create a set of possible norms. By doing this, we give the same probability to all goals of becoming hindered by a norm. Both punishments and rewards, in each norm, are also randomly generated. Then, a set of goals in conflict is created from the base of goals. Conflicts are represented as pairs of goals with the same probability of appearing in an experiment. Once the sets of possible goals, norms and conflicts are fixed, they are used until the end of the experiment as follows. Random subsets of goals and norms are extracted to represent the current norms and goals of each agent. In this way, norms are evaluated by agents following different strategies for norm compliance, so that similar inputs produce different outcomes.

5.2 Results

In our initial experimentation, we recorded both the *social contribution* and the *individual satisfaction* of each agent for a particular percentage of conflicts over a period of time. First no conflicts were considered, meaning that none of the norms conflicted with any goal. Then the experiment was repeated with the number of conflicts increased in a proportion of 25% until all norms conflict with goals. An experiment was run for each number of conflicts. Each test consisted of 100 runs. In each run, 10 goals and norms were used as base, and subsets of 5 norms and goals were taken randomly to represent the active norms and goals that each agent evaluates according to its selected strategies. The results of this experiment are illustrated in Figure 1, where each graph represents the normative behaviour of the corresponding agent. The vertical axis shows the values of both the *social contribution* (which graph is indicated by the squares) and the *individual satisfaction* of agents (which graph is indicated by the triangles). In this axis, the value 1 represents that one agent either provides the maximum social contribution (i.e. it complies with all norms) or achieves the maximum

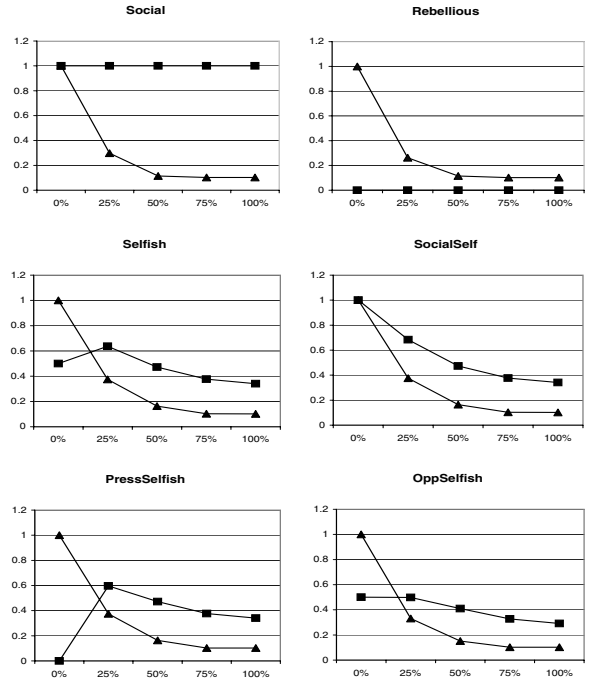

Figure 1: Normative Agent Behaviour

individual satisfaction (i.e. none of its goals is hindered by either normative goals or punishments). The horizontal axis shows the percentage of conflicts taken in each test.

As can be observed, in general, social strategies make societies very stable because social goals, expressed through norms, are almost guaranteed. However, individual satisfaction of this kind of agents decreases at the same rate as conflicts between normative goals and individuals goals. By contrast, rebellious individuals never contribute to social concerns even though their own satisfaction is not necessarily achieved, especially when punishments are applied. Selfish strategies, (those complying with norms only when either some of an agent's interests can be damaged, or some benefits can be obtained), scarcely provide social contributions. One important thing to observe here is that despite their selfishness, individual satisfaction is not guaranteed when conflicts between goals and norms increase, and of course when punishments are applied. By contrast, agents following social selfish strategies perform reasonably well in terms of their society contribution only when norms do not conflict with their goals.

6. CONCLUSIONS

By providing this formal approach to norm compliance we have addressed some of the aspects that must be considered in order to incorporate norms in agents. However, instead of taking a static view of norms in which norms are complied with straighforwardly, we adopt a more dynamic view in which an agent's motivations, and therefore its autonomy, play an important role. Our norm compliance model incorporates punishments and rewards as mechanisms for enforcing or encouraging the compliance with norms. However, contrary to other models, they are only taken into account if they hinder or benefit the goals of an

agent. The model itself was inspired by different research on norms where agents have the freedom to decide whether to comply or not with a norm [1, 6, 9]. However, our model provides a more general structure in which the particular cases of norm compliance that have been considered until now can be easily incorporated into an agent architecture. For example, Conte and Castelfranchi [6] compare two kind of agents: *incentive-based rational deciders* and *normative agents*, where the first comply with norms only if the utility of obedience is higher than the utility of transgression, and normative agents, by contrast, always fulfill them. As can be seen, both of these agents can be easily implemented in our model as *selfish* and *social* agents respectively. The authors also claim that a society composed of selfish individuals declines quickly and sometimes collapses in relation to the unsatisfied norms, whereas societies including normative agents always achieve their social goals. However, their model is intuitive, without either formalisation or experimentation to demonstrate the validity of the hypothesis, unlike the work described in this paper.

Similarly, Dignum et al. [9] describe a model of BDI agents in which obligations are fulfilled only if the cost of the punishment is higher than the cost of compliance. This particular view can be reduced to the simple case in our model of agents following a *pressured* strategy. In this way, besides offering a more complete model of autonomous norm compliance, our work additionally provides an analysis of the effects of norm compliance on both societies and individuals. Future work will provide more detailed analyses of the effects of different strategies in different societies, and how different views of norm adoption complicate this further.

Acknowledgments: The first author is supported by the Faculty Enhancement Program (PROMEP) of the Mexican Ministry of Public Education(SEP) and the Benemérita Universidad Autónoma de Puebla, México.

7. REFERENCES

[1] G. Boella and L. Lesmo. Deliberative normative agents. In Dellarocas and Conte [8].

[2] C. Castelfranchi, R. Conte, and M. Paolucci. Normative reputation and the cost of compliance. *Journal of Artificial Societies and Social Simulation*, 1(3), 1998.

[3] R. Conte. Emergent (info)institutions. *Journal of Cognitive Systems Research*, 2:97–110, 2001.

[4] R. Conte and C. Castelfranchi. *Cognitive and Social Action*. UCL Press, England, 1995.

[5] R. Conte and C. Castelfranchi. Norms as mental objects. From normative beliefs to normative goals. In C. Castelfranchi and J. P. Müller, editors, *From Reaction to Cognition (MAAMAW'93)*, LNAI 957, pages 186–196. Springer-Verlag, 1995.

[6] R. Conte and C. Castelfranchi. Are incentives good enough to achieve (info)social order? In Dellarocas and Conte [8].

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

[8] C. Dellarocas and R. Conte, editors. *Proceedings of the Workshop on Norms and Institutions in MAS (at AGENTS2000)*, Barcelona, Spain, 2000.

[9] F. Dignum, D. Morley, E. Sonenberg, and L. Cavendon. Towards socially sophisticated BDI agents. In *Proceedings of the Fourth International Conference on MultiAgent Systems (ICMAS2000)*, pages 111–118, Boston, USA, 2000. IEEE Computer Society.

[10] M. d'Inverno and M. Luck. *Understanding Agent Systems*. Springer-Verlag, 2001.

[11] S. Kerr. On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, 18(4):769–782, 1975.

[12] S. Kirn and L. Gasser. Organizational approaches to coordination in multi-agent systems. Technical report, IImenau Technical University, Germany, 1998.

[13] V. Lesser and L. Gasser, editors. *Proceedings of the First International Conference on MultiAgent Systems (ICMAS95)*, Sn. Francisco California, 1995. AAAI Press/MIT Press.

[14] M. Luck and M. d'Inverno. A formal framework for agency and autonomy. In Lesser and Gasser [13], pages 254–260.

[15] M. Luck and M. d'Inverno. Motivated behaviour for goal adoption. In C. Zhang and D. Lukose, editors, *Multi-Agents Systems. Theories Languages and Applications*, LNAI 1544, pages 58–73. Springer-Verlag, 1998.

[16] M. Luck and M. d'Inverno. Plan analysis for autonomous sociological agents. In Y. Lesperance and C. Castelfranchi, editors, *Intelligent Agents VII (ATAL00)*, pages 172–186, USA, 2000.

[17] M. Luck and M. d'Inverno. A conceptual framework for agent definition and development. *The Computer Journal*, 44(1):1–20, 2001.

[18] M. Macy. Social order in artificial worlds. *Journal of Artificial Societies and Social Simulation*, 1(1), 1998.

[19] A. Ross. *Directives and Norms*. Routledge and Kegan Paul Ltd., England, 1968.

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

[21] R. Tuomela and M. Bonnevier-Toumela. Social norms, task, and roles. Technical report, University of Helsinki, Helsinki, 1992.

[22] R. Tuomela and M. Bonnevier-Toumela. Norms and agreement. *European Journal of Law, Philosophy and Computer Sience*, 5:41–46, 1995.

[23] A. Walker and M. Wooldridge. Understanding the emergence of conventions in multi-agent systems. In Lesser and Gasser [13], pages 384–389.

[24] R. Wieringa, F. Dignum, J. Meyer, and R. Kuiper. A modal approach to intentions, commitments and obligations: Intention plus commitment yields obligation. In M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems (Workshops in Computing)*, pages 80–97. Springer-Verlag, 1996.

[25] F. Zambonelli, N. Jennings, and M. Wooldridge. Organisational abstractions for the analysis and design of multi-agent systems. In *Proceedings of the First International Workshop on Agent-Oriented Software Engineering*, 2000.