Determining the Willingness to Comply With Norms (Extended Abstract)

N. Criado
Universidad Politécnica de Valencia
Valencia (Spain)
ncriado@dsic.upv.es

E. Argente
Universidad Politécnica de Valencia
Valencia (Spain)
eargente@dsic.upv.es

P. Noriega
IIIA-CSIC
Campus de la UAB, Bellaterra, Catalonia (Spain)
pablo@iiia.csic.es

V. Botti
Universidad Politécnica de Valencia
Valencia (Spain)
vbotti@dsic.upv.es

ABSTRACT

In this paper, we propose that agents make decisions about norm compliance based on three different factors: self-interest, enforcement mechanisms and internalised emotions. Different agent personalities can be defined according to the importance given to each factor.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents

General Terms

Legal Aspects, Algorithms, Experimentation

Keywords

Norm Compliance, Norms, BDI Agents

1. INTRODUCTION

Despite the efforts that have been made to develop agents endowed with capabilities for taking into account norms in their decisions, the development of procedures for making autonomous decisions about norm compliance is an important issue that requires more attention [2].

Proposals on normative agent architectures can be mainly classified into *norm-oriented* and *goal-oriented*. The behaviour of norm-oriented agents is completely determined by norms and they do not make decisions about norm compliance. The behaviour of goal-oriented agents is determined by both norms and goals. Up to now the decisions about norm compliance consider the impact of norms and their enforcement mechanisms (i.e., sanctions and rewards) on the agents' goals. Obviously, these reasons are relevant for making decisions about norm compliance. However, there are works on the psychology field [3] that claim that norm compliance is not only explained by rational reasons that consider the impact of norms and their enforcement procedures (sanctions and rewards) on the agent's goals. Besides that, there are emotional reasons, which are related to emotions such as shame, that have not been considered yet in the development of norm-autonomous agents. In this paper we analyse how agents can determine their willingness to comply with norms according to rational and emotional factors.

2. DETERMINING THE WILLINGNESS TO COMPLY WITH NORMS

As stated by Conte et al. in [1] "*the decision to comply with a norm is made considering: the value of the violation (probability and weight of punishment), the importance of the goal and feelings related to norm violation*". To calculate this willingness we have mainly considered the works of Elster [3] that analyse factors that sustain norms in human societies. In these works, Elster claims that compliance with norms can be explained by three factors: (i) *self-interest* motivations (f'_w), which consider the influence of norm compliance and violation on agent's goals; (ii) the *expectations* (f''_w) of being rewarded or sanctioned by others; and (iii) *emotional* factors (f'''_w) that are related to internalised emotions such as honour (vs. shame) and hope (vs. fear). The agent's willingness to follow a concrete norm is calculated as a weighted average as follows:

$$\frac{w' \times f'_w + w'' \times f''_w + w''' \times f'''_w}{w' + w'' + w'''}$$

where the weights w', w'' and w''' are defined within the $[0, 1]$ interval.

We have assumed that the weighted average is a suitable method to derive the central tendency of these three functions. The weights that each agent gives to these factors characterise the agent's personality and do not depend on the norm that is considered.

2.1 Self-interest

The *self-interest* factor (f'_w) evaluates the consequences of a given norm from a utilitarian perspective; i.e., the utility is the good to be maximized. The utility of a norm is defined

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

by considering the direct positive or negative consequence of the norm fulfilment. In case of an obligation, the direct consequence of the fulfilment of the obligation is the obliged condition. In case of a prohibition, obeying this prohibition implies that the forbidden condition will be avoided.

2.2 Expectations

The *expectation* factor (f''_w) models the impact of the external enforcement on agents. Specifically, the enforcement mechanism considered in this work consists in a material system of sanctions and rewards that modify the utility that agents obtain when they violate or fulfil norms. This factor considers how much the agent loses from being penalised and how much it gains from being rewarded. The violation of the norm implies that the agent will be sanctioned and not rewarded. Thus, the *expectation* factor is defined as the combination of the undesirability of the sanction and the negation of the reward. For simplicity, we assume that there is a perfect enforcement that always punishes offenders and rewards obedience. However, if agents are able to perceive the probability of being punished or rewarded, then the desirability of sanctions and rewards should be pondered with these probabilities.

2.3 Anticipated Emotions

The *emotional* factor (f'''_w) models the emotions triggered when the agent violates a given norm. We use the term emotion for representing the valued reaction of agents (i.e., the agent's cognitive interpretation) with respect to some aspect of the world (i.e., the reality) [4]. Specifically, agents are capable of anticipating, exhibiting and explaining those human emotions that are involved with the normative decisions. Thereby, the decisions about norm compliance are based on other criteria beyond utility.

As argued by Elster in [3], in humans norms are sustained by the desire to avoid the disapproval of others. Following Elster's proposal, when the violation of norms is greeted with condemnation self-*attribution* emotions (i.e., shame) are triggered on the offender. Moreover, the situations that are predicted to occur when norms are violated may cause *prospect* emotions (i.e., hope and fear) on the offender.

To estimate the value of these two emotions an emotional model susceptible of being implemented in a software agent is required. Specifically, we consider one of the emotional models that have made a deeper impact on the *Multi-Agent System* (MAS) field; the OCC model developed by *Ortony, Clore and Collins* in [4]. Thus, the OCC model has been used for establishing the intensity of the emotions that are involved in the norm-reasoning process as follows:

- *Self-Attribution Emotions.* According to the OCC model, shame is a self-attribution emotion that is elicited by the evaluation of the actions that have been performed by the agent itself. Specifically, the shame that the agent will feel if it violates a given norm is defined by considering the salience of this norm. Therefore, self-attribution emotions only sustain norm obedience.

- *Prospect Emotions.* According to the OCC model, the hope (vs. fear) emotion is triggered when a desirable (vs. undesirable) event is predicted. The fear and hope emotions that may be triggered if a norm is violated are defined by considering the desirability and probability of the consequences of violating and norm.

3. MAIN AGENT TYPES

The decisions about norm compliance are made by considering three willingness factors (i.e., f'_w, f''_w and f'''_w) that are combined as a weighted average. Therefore, different agent personalities can be modelled according to the definition of the weights w', w'' and w'''. The three basic personalities are:

- *Egoist agents* ($w' = 1$, $w'' = 0$ and $w''' = 0$) are the lest prone to comply with norms, since they only consider whether the norm condition favours or hinders their goals.

- *Cautious agents* ($w' = 0$, $w'' = 1$ and $w''' = 0$) are more prone to comply with norms than egoist agents. This can be explained by the fact that cautious agents consider whether either the sanction or the negation of the reward favour their goals.

- *Emotional agents* ($w' = 0$, $w'' = 0$ and $w''' = 1$) are the most willing to obey norms; i.e., they are the most norm-oriented. This is explained by the fact that attribution emotion only sustains norm obedience.

4. CONCLUSIONS

This paper answers a main question that is related to the possibility of developing norm-autonomous agents that consider emotional criteria in their decisions about norm compliance. In response to this issue, this paper describes how agents can consider both their preferences and the norm repercussions when they determine their willingness to comply with norms. The repercussion of norms is not only defined in terms of the utility of norms and the economic cost (vs. benefit) of the sanctions (vs. rewards), but also in terms of the social repercussion of norms (i.e., emotional factors). Specifically, agents are endowed with mechanisms for anticipating the emotions that will be elicited if the norms are transgressed. Moreover, the way in which agents combine rational and emotional factors allow different personalities to be modelled. As future work we plan to evaluate whether the use of agents that consider emotional criteria obtains better results in norm-governed scenarios.

5. ACKNOWLEDGMENTS

This paper was partially funded by the Spanish government under grants CONSOLIDER-INGENIO 2010 CSD2007-00022, TIN2009-13839-C03-01, TIN2008-04446 and by the FPU grant AP-2007-01256 awarded to N. Criado. This research has also been partially funded by the Generalitat de Catalunya under the grant 2009-SGR-1434 and Valencian Prometeo project 2008/051.

6. REFERENCES

[1] R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm acceptance. In *Proc. of ATAL*, pages 99–112, 1999.

[2] N. Criado, E. Argente, and V. Botti. Open Issues for Normative Multi-Agent Systems.

[3] J. Elster. Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117, 1989.

[4] A. Ortony, G. Clore, and A. Collins. *The cognitive structure of emotions.* Cambridge Univ Pr, 1990.