Self-Organization in Decentralized Agent Societies through Social Norms

(Extended Abstract)

Daniel Villatoro
Artificial Intelligence Research Institute (IIIA) - Spanish Scientific Research Council (CSIC)
Bellatera, Barcelona, Spain
dvillatoro@iiia.csic.es

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems

General Terms

Experimentation

Keywords

Artificial social systems, Social and organizational structure, Self-organisation, Norms

1. INTRODUCTION

Social norms help people self-organizing in many situations where having an authority representative is not feasible. On the contrary to institutional rules, the responsibility to enforce social norms is not the task of a central authority but a task of each member of the society. *"The social norms I am talking about are not the formal, prescriptive or proscriptive rules designed, imposed, and enforced by an exogenous authority through the administration of selective incentives. I rather discuss informal norms that emerge through the decentralized interaction of agents within a collective and are not imposed or designed by an authority"*[3]. In recent years, the use of social norms has been considered also as a mechanism to regulate virtual societies and specifically heterogeneous societies formed by humans and artificial agents.

One of the main topics of research regarding the use of social norms in virtual societies is how they emerge, that is, how social norms are created at first instance. We divide the emergence of norms into two different stages: (a) how norms appear in the mind of one or several individuals and (b) how these new norms are spread over the society until they become accepted social norms. We are interested in studying the second stage, the spreading and acceptance of social norms, what Axelrod [2] defines as *norm support*. Our understanding of norm support deals with the problem of which norm is established as the dominant. Specifically, we deal with two different branches of the research on normative sytems: conventional norms and essential norms. As described in [6], on the one hand conventional norms fix one norm amongst a set of norms that are equally efficient as long as every member of the population uses the same (e.g. communication protocols, greetings, driving side of the road), and on the other hand, essential norms solve or ease collective action problems, where there is a conflict between the individual and the collective interests. The scientific question of this research is how to accelerate the establishment of a common norm in virtual societies: in the case of conventional norms, by dissolving the subconventions; and in the case of essential norms, by studying the effects of punishment and norm internalization.

Cite as: Self-Organization in Decentralized Agent Societies through Social Norms (Extended Abstract), Daniel Villatoro, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 1373-1374.

2. CONVENTIONAL NORMS

The social topology that restricts agent interactions plays a crucial role on any emergent phenomena resulting from those interactions [4]. *Convention emergence* is one mechanism for sustaining social order, increasing the predictability of behavior in the society and specify the details of those unwritten laws. Examples of conventions pertinent to MAS would be the selection of a coordination protocol, communication language, or (in a multitask scenario) the selection of the problem to be solved. Conventions help agents to choose a solution from a search space where potentially all solutions are equally good, as long as all agents use the same.

In *social learning* [5] of norms, where each agent is learning concurrently over repeated interactions with randomly selected neighbours in the social network, a key factor influencing success of an individual is how it learns from the "appropriate" agents in their social network. Therefore, agents can develop subconventions depending on their position on the topology of interaction. As identified by several authors, metastable subconventions interfere with the speed of the emergence of more general conventions. The problem of subconventions is a critical bottleneck that can derail emergence of conventions in agent societies and mechanisms need to be developed that can alleviate this problem. Subconventions are conventions adopted by a subset of agents in a social network who have converged to a different convention than the majority of the population.

Subconventions are facilitated by the topological configuration of the environment (isolated areas of the graph which promote endogamy) or by the agent reward function (concordance with previous history, promoting cultural maintenance). Assuming that agents cannot modify their own reward functions, the problem of subconventions has to be solved through the topological reconfiguration of the envi-

ronment.

Agents can exercise certain control over their social network so as to improve one's own utility or social status. We define *Social Instruments* to be a set of tools available to agents to be used within a society to influence, directly or indirectly, the behaviour of its members by exploiting the structure of the social network. Social instruments are used independently (an agent do not need any other agent to use a social instrument) and have an aggregated global effect (the more agents use the social instrument, the stronger the effect).

3. ESSENTIAL NORMS

The problem social scientists still revolve around is how autonomous systems, like living beings, perform positive behaviors toward one another and comply with existing norms, especially since self-regarding agents are much better-off than other-regarding agents at within-group competition. Since Durkheim, the key to solving the puzzle is found in the theory of internalization of norms. One plausible explanation of voluntary non self-interested compliance with social norms is that norms have been internalized.

Internalization occurs when *"a norm's maintenance has become independent of external outcomes - that is, to the extent that its reinforcing consequences are internally mediated, without the support of external events such as rewards or punishment"* [1, p 18].

Agents conform to an internal norm because so doing is an *end* in itself, and not merely because of external sanctions, such as material rewards or punishment. This internalization process will not only benefit agents for the actual norm compliance, but will also benefit the society as a whole by reducing the actual costs of norm enforcement. Despite these important contributions, however, the community's scientific definition and understanding of the process of norm internalization is still fragmentary and insufficient.

The main purpose of our research is to argue for the necessity of a rich cognitive model of norm internalization in order to (a) provide a unifying view of the phenomenon, accounting for the features it shares with related phenomena (e.g., robust conformity as in automatic behavior) and the specific properties that keep it distinct from them (autonomy); (b) model the process of internalization, i.e. its proximate causes (as compared to the distal, evolutionary ones, like in the work of Gintis); (c) characterize it as a progressive process, occurring at various levels of depth and giving rise to more or less robust compliance; and finally (d) allow for flexible conformity, enabling agents to retrieve full control over those norms which have been converted into automatic behavioral responses.

Thanks to such a model of norm internalization, it has been possible to adapt existing agent architectures (EMIL-A evolved to EMIL-I-A) and to design a simulation platform to test and answer a number of hypotheses and questions such as: Which types of mental properties and ingredients ought individuals to possess in order to exhibit different forms of compliance? How sensitive each modality is to external sanctions? What are the most effective norm enforcement mechanisms? How many people have to internalize a norm in order for it to spread and remain stable? What are the different implications for society and governance of different modalities of norm compliance?

This cognitive architecture have also helped us explore the specific ways in which punishment and sanction favor the achievement of cooperation and the spreading of social norms in social systems populated by autonomous agents. Because of the similarity between punishment and sanction, these two phenomena are often mistaken one for another and considered as a *single* behavior. We claim that punishment and sanction are different behaviours and that can be distinguished on the basis of their mental antecedents and of the way in which they aim to influence the future conduct of others.

On the one hand, punishment is a practice consisting in imposing a fine to the wrongdoer, with the aim of deterring him from future offenses. Deterrence is achieved by modifying the relative costs and benefits of the situation, so that wrongdoing turns into a less attractive option. The effect of punishment is achieved by increasing individuals' expectations about the price of non-compliance. This view of punishment is in line with the one supposed by the Beckerian economic model of crime and with the approach adopted by experimental economics. On the other hand, sanction works by imposing a cost, as punishment does, and in addition by *communicating* to the target (and possibly to the audience) both the existence and the violation of a norm. The sanctioner ideally wants to induce the agent to comply with the norm not just to avoid punishment, but because he recognizes that there is a norm and wants to observe it for its own sake.

We argue that norm compliance will be more robust if agents are enforced by sanction: where people have internal motivations to follow the norms, the frequency of compliance in the population will be higher than if people observe the norm only instrumentally (when it is in their interest to do so). Sanction are powerful social tools allowing norms and institution to be viable and robust across time.

4. ACKNOWLEDGMENTS

Daniel Villatoro is supported by a CSIC predoctoral fellowship under JAE program. Daniel Villatoro also thanks Sandip Sen, Rosaria Conte, Giulia Andrighetto and specially Jordi Sabater-Mir for their wise advices.

5. REFERENCES

[1] J. M. Aronfreed. *Conduct and conscience; the socialization of internalized control over behavior [by] Justin Aronfreed*. Academic Press, New York,, 1968.

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

[3] C. Bicchieri. *The Grammar of Society: The nature and Dynamics of Social Norms*. Cambridge University Press, 2006.

[4] J. E. Kittock. The impact of locality and authority on emergent conventions: initial observations. In *Proceedings of AAAI'94*, volume 1, pages 420–425. American Association for Artificial Intelligence, 1994.

[5] S. Sen and S. Airiau. Emergence of norms through social learning. *Proceedings of IJCAI-07*, pages 1507–1512, 2007.

[6] D. Villatoro, S. Sen, and J. Sabater-Mir. Of social norms and sanctioning: A game theoretical overview. *International Journal of Agent Technologies and Systems*, 2:1–15, 2010.