Social Norms for Self-Policing Multi-Agent Systems and Virtual Societies* (Extended Abstract)

Daniel Villatoro
Artificial Intelligence Research Institute (IIIA) -
Spanish Scientific Research Council (CSIC), Bellaterra, Spain
Barcelona Digital Technology Centre, Barcelona, Spain
dvillatoro@bdigital.org

Abstract

Social norms are one of the mechanisms for decentralized societies to achieve coordination amongst individuals. Such norms are conflict resolution strategies that develop from the population interactions instead of a centralized entity dictating agent protocol. One of the most important characteristics of social norms is that they are imposed by the members of the society, and they are responsible for the fulfillment and defense of these norms. By allowing agents to manage (impose, abide by and defend) social norms, societies achieve a higher degree of freedom by lacking the necessity of authorities supervising all the interactions amongst agents. In this article we summarize the contributions of my dissertation, where we provide an unifying framework for the analysis of social norms in virtual societies, providing an strong emphasis on virtual agents and humans.

1 Introduction

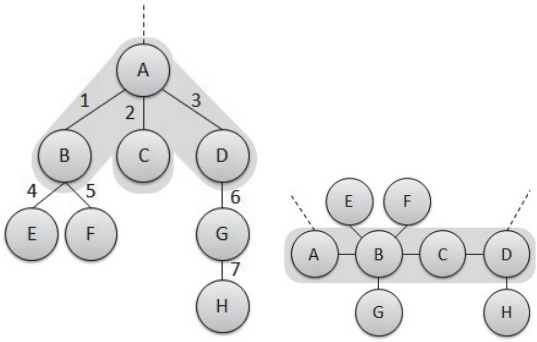
Social norms are part of our everyday life. They help people self-organizing in many situations where having an authority representative is not feasible. On the contrary to institutional rules, the responsibility to enforce social norms is not the task of a central authority but a task of each member of the society. From the book of Bicchieri [Bicchieri, 2006], the following definition of social norms is extracted: *"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 col- lective and are not imposed or designed by an authority"*.

Social norms are used in human societies as a mechanism to improve the behaviour of the individuals in those societies without relying on a centralized and omnipresent authority. In recent years, the use of these kinds of norms has been considered also as a mechanism to regulate virtual societies and specifically heterogeneous societies formed by humans and artificial agents. From another point of view, the possibility of performing agent based simulation on social norms helps us to understand better how they work in human societies. 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. This has been studied by several authors who propose different factors that can influence this emergence. 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 [Axelrod, 1986] defines as norm support. Our understanding of norm support deals with the problem of which norm is established as the dominant. Specifically, this dissertation [Villatoro, 2011] deals with two different branches of the research on normative sytems: conventional norms and essential norms. As described elsewhere [Villatoro *et al.*, 2010], 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.

2 Conventional Norms

The social topology that restricts agent interactions plays a crucial role on any emergent phenomena resulting from those interactions [Kittock, 1994]. 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.

*The dissertation on which this extended abstract is based was the recipient of the IFAAMAS-11 Victor Lesser Distinguished Dissertation Award [Villatoro, 2011].



(a) Claw (b) Caterpillar

Figure 1: Self-Reinforcing Structures

In social learning [Sen and Airiau, 2007] 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. One of the most important contributions of this research is identifying that 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), as identified in [Villatoro *et al.*, 2011d]. Specifically, we have discovered that certain type of topological configurations (mainly identified in social networks, and shown in Fig.1) posses certain type of self-reinforcing substructures that facilitates the emergence of meta-stable subconventions.

Assuming that agents cannot modify their own reward functions, the problem of subconventions has to be solved through the topological reconfiguration of the environment. Agents can exercise certain control over their social network so as to improve one's own utility or social status. We define Social Instruments [Villatoro *et al.*, 2011b] 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).

The identification of the Self-Reinforcing Structures have allowed us to develop the necessary mechanisms to reach full convergence, which was never previously reached by any other researcher in the community [Villatoro *et al.*, 2011c].

3 Essential Norms

Previously we have seen how conventions can be delayed by the emergence and maintenance of subconventions. As analyzed, subconventions are mainly generated by the topological configuration of the interaction network. However, this situation emerges as we dealt with the most strict and pure definition of convention, where all the options are potentially equally good as long as the whole population follows the same convention. Consequently it is in the agents' self-interest to accept the convention with which they obtain maximum benefit. However, empirical results made us learnt that certain agents obtain a benefit from following a convention, but due to some exogenous reasons they can be positioned in a frontier. Being located in a frontier forces an agent to maintain the convention with which the highest benefit is obtained; because of the social learning approach, agents obtain a larger benefit from the convention followed by the majority of their neighbours. The topological configuration can produce the neighbours not to change their convention so our frontier agent interacts successfully with all its neighbours. Because of that reason, some of the interactions of the frontier agent (depending on the number of uncoordinated neighbours) will be unsuccessful. At that moment, all the agents related to the agent(s) in the frontier are interested in reaching a common convention (and resolve the subconvention), to reduce the number of unsuccessful interactions, and maximize their utility.

Agents' strategy are ruled by a utility-maximizing policy, as classically though for human decision making [Becker, 1968]. Therefore, by providing agents with the correct mechanisms, they would be able to reduce others utility depending on their behaviour. This type of action is commonly known as a punishment or a sanction[1]. Theoretical, empirical and ethnographic studies have demonstrated that punishment in human societies promotes and sustains cooperation in large groups of unrelated individuals and more generally plays a crucial role in the maintenance of social order [Fehr and Gachter, 2002; Sigmund, 2007]. Normally, punishment implies a cost for both the punisher and the punished, reducing both utilities. However and because of these costs associated to punishment, the unilateral decision of punishment from one single agent to another conveys a second-order social dilemma [Dreber *et al.*, 2008]. We hypothesized that by dividing the costs of punishment amongst the members of the society, this second-order public good becomes less costly, and therefore, more attractive for the members of the society: at a small costs, the potential benefits are increased as the number of freeriders decrease. In that way, we can see how a group of agents could coordinate for punishing a certain agent, whose behaviour should be changed for the benefit of the society, in our case, the frontier agent.

3.1 Solving the Subconvemtions through Distributed Punishment

Within the scope of the MacNorms project [MacNorms, 2008], and in conjunction with the *Instituto de Análisis*

[1]We will see later the difference between punishment and sanction at a cognitive level.

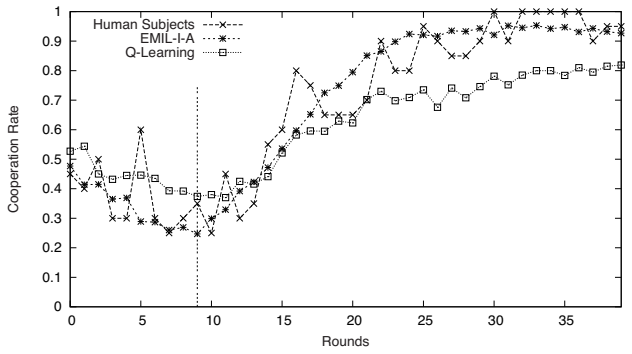


Figure 2: Results from the Distributed Punishment Experiment.

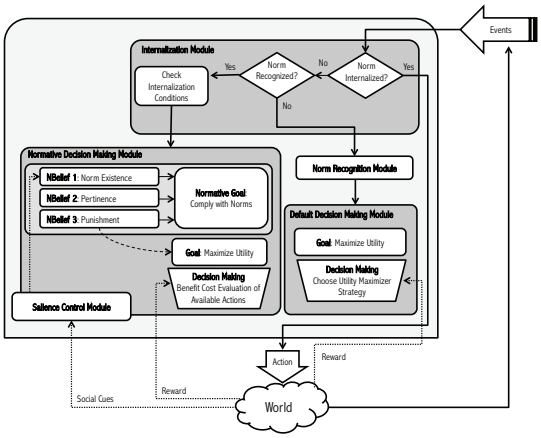


Figure 3: EMIL-I-A Architectural Design

Económico, we performed experiments with human subjects using the HIHEREI platform [Brito *et al.*, 2009]; empirical findings proved that Distributed Punishment was effective (in comparison with classical Monolateral punishment) when applied although difficult to achieve without the possibility of communication. Moreover, these results led us to put forward the hypothesis that with the same material incentive inflicted, distributed punishment is more effective in enhancing compliance than mono-lateral one, because it is more likely to be interpreted as a sanction, even when the educational message is implicit (i.e., it is not conveyed through verbal communication). The reason is because the higher the number of punishers, the less likely the observers will interpret their punishing behaviors as dictated by the self-interest and, conversely, the more likely they will attribute punishment to impersonal, normative and possibly legitimate reasons, i.e. as upholding a norm.

To test it, a laboratory experiment populated both by humans and virtual agents and reproducing a social dilemma scenario has been conducted. The experiment conducted in the laboratory has also been replicated by agent-based simulation, obtaining convergent results (as shown in Fig. 2). These data provide support for the hypothesis that punishment is effective in regulating people's behavior not only through economic incentives, but also thanks to the normative information it conveys and the normative request it asks of people [Villatoro *et al.*, 2012]. The comparison of the results from the experiments with humans and from the simulations shows that the simulation model captures the essential features of the human data. The simulation model is not intended just to replicate in silico the experimental findings, but it is an attempt to provide an explicit model of the cognitive mechanisms and processes allowing distributed punishment to positively promote compliant conduct. By comparing the performance of our proposed cogno-normative architecture with other classical reinforcement learning architectures, we have observed that the former reproduces behavioral dynamics more similar to humans than the latter ones.

3.2 The Normative Power of Sanctions

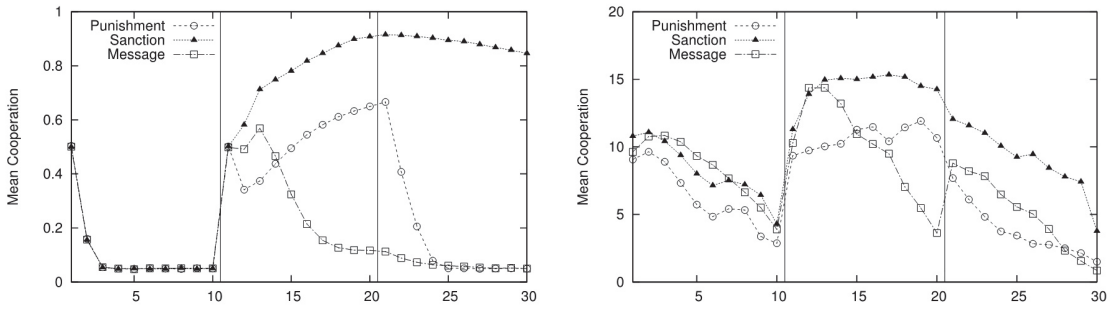
Based on this empirical result, and those obtained by others [Noussair and Tucker, 2005], we hypothesize that some behaviours (like a consistent punishment from the entire group) might contain implicit normative messages that affect to the decision making in a more profound way than a mere benefit-cost calculation. *Human subjects do not only take into consideration the potential punishment that can be received, but they rather interpret some social cues as the existence of a social norm, that even in the absence of punishment systems, can lead the subject to abide by the norms*.

As far as we are concerned, there is no existing agent architecture that is able to identify these type of cues and uses them into its decision making. Consequently, we have envisioned a cognitive architecture that incorporate a more complex reasoning than a simple benefit-cost calculation.

This architecture, EMIL-I-A (represented in Fig. 3) consists of mechanisms and mental representations allowing norms to affect the behavior of autonomous intelligent agents [Andrighetto *et al.*, 2010]. As any BDI-type (Belief-Desire and Intention) architecture EMIL-I-A operates through modules for different sub-tasks (recognition, adoption, decision making, salience control, etc...) and acts on mental representations for goals and beliefs in a non-rigid sequence.

EMIL-I-A adopts a definition of norms in which social norms are not static objects and the degree to which they are operative and active varies from group to group and from one agent to another: we refer to the degree of activation as norm's *salience*. The more salient a norm is, the more it will elicit a normative behaviour [Bicchieri, 2006; Xiao and Houser, 2005; Cialdini *et al.*, 1990]. Norm's salience is a complex function, depending on several social and individual factors allowing agents to *dynamically* monitor if the normative scene is changing and to adapt to it[2]. This flexible notation of norm makes our normative agents as autonomous as socially responsive. They are autonomous in that they act on their own beliefs and goals (on the basis of

[2]It is interesting to notice that this mechanism allows agents to record the social and normative information, without necessarily proactively exploring the world (e.g. with a trial and error procedure).



(a) Humans' Experimental Results (b) EMIL-I-As' Experimental Results

Figure 4: Effects of Punishment, Sanction and Message on Cooperation Rates.

their salience). However, they are also responsive to their environment, and to the inputs they receive from it, especially to social inputs.

Based on our results with human subjects in the Distributed Punishment Experiment, and under the suspicion that different types of punishment have a different effect on the emergence of cooperation, we performed new experiments with human subjects that made explicit the difference between punishment (cost imposition) and sanction(cost imposition plus normative ellicitation), as suggested in [Giardini *et al.*, 2010]. Experimental results are shown in Fig. 4(a). As far as we know, this is the first time this result has been proven, being therefore interesting for such community. This results not only allowed us to confirm such difference in which EMIL-I-A is based, but also to fine-tune the specific parameters that set up the Norm Salience Control module. On the other hand, EMIL-I-A has also performed well recreating the human subjects dynamics in the same experiment, as shown in Fig. 4(b).

After proving the validity of such architecture, we (computer scientists and policy makers) are given with a powerful tool that allows us to further exploit these punishment technologies, in different situations, unfeasible to obtain in the laboratory with human subjects. The simulation results obtained by exploiting the capabilities of EMIL-I-A show the ways in which punishment and sanction affects the emergence of cooperation [Villatoro *et al.*, 2011a]. More specifically, these results seem to verify our hypotheses that the signaling component of sanction allows this mechanism (a) to be more effective in the achievement of cooperation; (b) to spread norm faster and wider in the population, making it more resilient to environmental change than if enforced only by mere punishment; (c) to reduce significantly the social cost for cooperation to emerge.

3.3 Internalization

Finally, we explore another cognitive mechanism that would allow us to explain the voluntary non self-interested compliance, Internalization.

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"[Aronfreed, 1968].

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 [Andrighetto *et al.*, 2010], it has been possible to adapt existing agent architectures (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? 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?

Acknowledgments

I would like to acknowledge my supervisor Dr. Jordi Sabater-Mir for the constant and helpful support he has provided me along the years of development of this dissertation. Also, I would like to acknowledge Dr. Giulia Andrighetto, as she has actively contributed with an important part of the research described in the dissertation.

References

[Andrighetto *et al.*, 2010] Giulia Andrighetto, Daniel Villatoro, and Rosaria Conte. Norm internalization in artificial societies. *AI Commun.*, 23(4):325–339, 2010.

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

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

[Becker, 1968] Gary S. Becker. Crime and punishment: An economic approach. *The Journal of Political Economy*, 76(2):169–217, 1968.

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

[Brito *et al.*, 2009] Ismel Brito, Isaac Pinyol, Daniel Villatoro, and Jordi Sabater-Mir. Hiherei: human interaction within hybrid environments regulated through electronic institutions. In *AAMAS '09: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 1417–1418, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems.

[Cialdini *et al.*, 1990] Robert B. Cialdini, Raymond R. Reno, and Carl A Kallgren. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015–1026, June 1990.

[Dreber *et al.*, 2008] Anna Dreber, David G. Rand, Drew Fudenberg, and Martin A. Nowak. Winners don't punish. *Nature*, 452(7185):348–351, March 2008.

[Fehr and Gachter, 2002] E Fehr and S Gachter. Altruistic punishment in humans. *Nature*, 415:137–140, 2002.

[Giardini *et al.*, 2010] F. Giardini, G. Andrighetto, and R. Conte. A cognitive model of punishment. In *COGSCI 2010, Annual Meeting of the Cognitive Science Society 11–14 August 2010,*. Portland, Oregon, 2010.

[Kittock, 1994] James 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.

[MacNorms, 2008] MacNorms. *MacNorms: Mechanisms for Self Organization and Social Control generators of Social Norms*. http://www.iiia.csic.es/es/project/macnorms-0, 2008.

[Noussair and Tucker, 2005] C.N. Noussair and S. Tucker. Combining monetary and social sanctions to promote cooperation. Open access publications from tilburg university, Tilburg University, 2005.

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

[Sigmund, 2007] K. Sigmund. Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology & Evolution*, 22(11):593–600, November 2007.

[Villatoro *et al.*, 2010] Daniel Villatoro, Sandip Sen, and Jordi Sabater-Mir. Of social norms and sanctioning: A game theoretical overview. *International Journal of Agent Technologies and Systems*, 2:1–15, 2010.

[Villatoro *et al.*, 2011a] Daniel Villatoro, Giulia Andrighetto, Jordi Sabater-Mir, and Rosaria Conte. Dynamic sanctioning for robust and cost-efficient norm compliance. *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 414–419, 2011.

[Villatoro *et al.*, 2011b] Daniel Villatoro, Jordi Sabater-Mir, and Sandip Sen. Social instruments for convention emergence. *AAMAS 2011, Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems.*, pages 1161–1162, 2011.

[Villatoro *et al.*, 2011c] Daniel Villatoro, Jordi Sabater-Mir, and Sandip Sen. Social instruments for robust convention emergence. *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 420–425, 2011.

[Villatoro *et al.*, 2011d] Daniel Villatoro, Sandip Sen, and Jordi Sabater-Mir. Exploring the dimensions of convention emergence in multiagent systems. *Advances in Complex Systems*, 14(2):201–227, 2011.

[Villatoro *et al.*, 2012] Daniel Villatoro, Giulia Andrighetto, Jordi Brandts, Jordi Sabater-Mir, and Rosaria Conte. Distributed punishment as a norm-signalling tool. pages 1189–1190, 2012.

[Villatoro, 2011] Daniel Villatoro. *Social norms for self-policing multi-agent systems and virtual societies*. PhD thesis, Universitat Autònoma de Barcelona, 2011.

[Xiao and Houser, 2005] E. Xiao and D. Houser. Emotion expression in human punishment behavior. *Proc Natl Acad Sci U S A*, 102(20):7398–7401, May 2005.