Of Social Norms and Sanctioning: A Game Theoretical Overview[1]

Daniel Villatoro, Artificial Intelligence Research Institute (IIIA - CSIC), Spain
Sandip Sen, University of Tulsa, USA
Jordi Sabater-Mir, Artificial Intelligence Research Institute (IIIA - CSIC), Spain

ABSTRACT

“Social norms” is a term widely used in different areas of research like sociology, philosophy or multiagent systems. However, there is still not a clear definition of what social norms are and the types of problems they solve. This work presents a general classification and distinction of norms from a game theoretical perspective. The type of norms treated in this work are those norms created through the interaction of agents and that are not imposed by any central authority. The main differentiation is made between *convetional norms* and *essential norms*. The former are norms created to establish a convention in a situation where several solutions are equally feasible, but the society must decide on one, e.g., driving on one side of the road; the later norms solve problems of collective action. Finally, we analyze several aspects of sanctioning mechanisms and how these mechanisms affect in the emergence of norms.

Keywords: Agent; Conventions; Game Theory; Sanctions; Social Norms

INTRODUCTION

Descriptions of tasks like greeting another person, dressing, driving, etc. are often accompanied by the phrase “in a proper way”. The “proper way” to fulfil these interaction protocols is specified by social norms. A number of tasks that require some kind of interaction with other agents might require agents to follow specified guidelines to successfully complete these tasks. Social norms can facilitate such agent interactions and enable agents to complete these tasks efficiently. Such social norms can emerge and spread among the society until they are widely accepted and adopted. Therefore, we can view social norms as key elements that enable coordination and self-organization in our everyday life.

Not all the social norms, however, deal with the same kind of interaction scenarios. We observe that social norms like greeting (shaking hands, kissing, leaning towards each other, or a simple “hi!”) pertain to different situations compared to, for example, the social norm of

[1] This version of the paper is not he final version of the paper. For that one, go to http://www.igi-global.com/articles/details.asp?ID=35764

recycling. We also observe that social norms, though referring to the same concept, are defined using different terms in the literature, e.g. norms, social laws, conventions, social norms.

In addition to the different types of norms, and the wide variety of terms used to define this social instrument, the study of social norms is made more challenging by the heterogeneous perspective on this issue and how it is viewed in diverse research areas such as economics, social sciences or multiagent systems. We believe that though these areas have interesting theories and practices to contribute to social normsliterature and can benefit from prudent adaptions and applications of social norms, not enough attention and effort has been expended on this potentially effective social coordination mechanism by the corresponding research groups.

The primary goal of this paper is to capture the different definitions and points of view of social norms from the related research areas and adapt them to a multiagent perspective. We also develop a characterization of social norms into two primary groups: coordination norms and essential norms. This division is also analyzed from a game-theoretical point of view with the goal of understanding the process of norm emergence. Finally, an analysis of the relation between social norms and sanctions[1] is presented.

NORMATIVE VOCABULARY

Before proceeding further we need to define some terms that are related to norms and that we consider to be the basic vocabulary for a common understanding of the three main branches of research (sociology, economy and multiagent systems). The interactionist norms that we are analyzing in this work are created, oriented, controlled and imposed by agents. Following Coleman (1998) agents are grouped by their role in the norm. There are two basic roles: the beneficiaries and the targets. *Targets* are the actors for whom the norm is specified for. *Beneficiaries* are those actors who benefit from the norm, potentially hold the norm and are potential sanctioners of the target actors. In the same example from Coleman, in the norm “Children should be seen and not heard”, the *target* are childrens and the *beneficiaries* are adults around those children looking for some peaceful environment.

Another characteristic of norms describes how the norms affect the actors.
The norms where the set of *target* and *beneficiaries* are completely disjoint are defined by Coleman as *Disjoint norms*.

However, the set of *target* actors and *beneficiaries* might not necessarily be disjoint for a norm. Coleman defines the norms where each actor is simultaneously beneficiary and target of the norm as *Conjoint norms*.
However these distinction are the extremes. Coleman presents different intermediate cases with different types of inclusions of both sets of targets and beneficiaries shown in Figure 1.

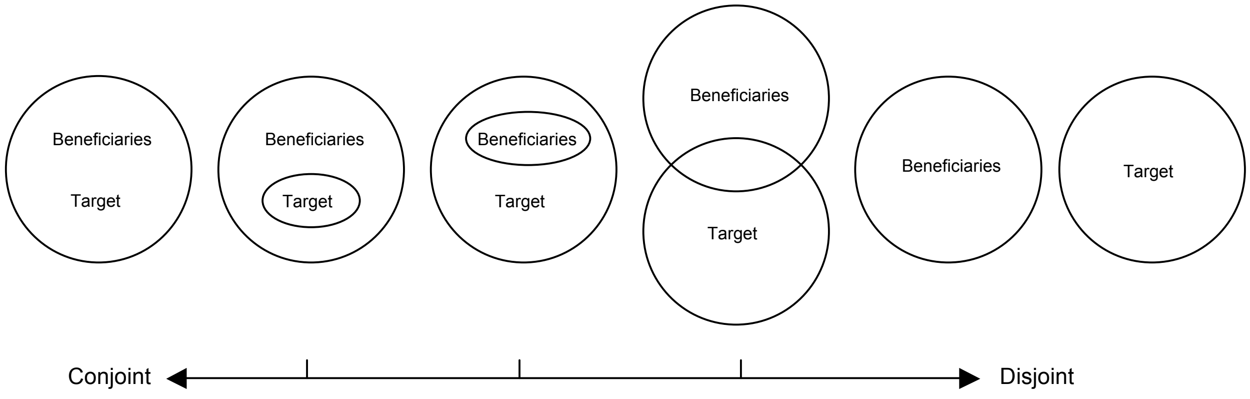

Figure 1. Coleman's inclusion relation of beneficiaries and targets of a norm for different types of norms

Coleman (1998) also claims that norms are directed at certain *focal actions*. The term *focal action* is directly borrowed from game theory where it exists the concept of *focal point*. A focal point is defined as a solution that players will tend to use in the absence of communication, because it seems natural, special or relevant to them. For example, imagine that you and your partner are visiting Paris. It is the first time for both of you in that city. Unfortunately, you are not in the same hotel and you have no means to communicate with each other, although you know that you have to meet each other at a certain time in a public place. You can choose between all the public places in Paris. Using a common sense reasoning, you might choose to go to the Eiffel Tower, the Pyramid of the Louvre Museum, or the Arc de Triomphe. Those would be focal points in the decision game. Consequently the focal action will be the action taken by the agents in the absence of communication.

Coleman also defines *externalities* as actions of individuals or collective actors which cause costs (negative externalities) or benefits (positive externalities) to other actors. Those who are affected by negative externalities have an interest in establishing norms that eliminate or reduce the externalities. Those who cause positive externalities have an interest in establishing a norm to be compensated. Boella and Lesmo (2002) discuss the existence of externalities and their role in the emergence of norms. Although they do not refer to them explicitly as norms, they affirm that "every action of an agent has an impact on the choices of other agents who can react to it".

Moreover, we present some characteristics that are common to all norms, and will help us later define the type of norm.

Coleman (1998) defines two characteristics of norms: norms can be either proscriptive or prescriptive.

Proscriptive norms are those that discourage or proscribe a focal action. In other words, *proscriptive norms* are those with the form "Do not ...". On the other hand, *prescriptive norms* encourage or prescribe a focal action. Similarly, these *prescriptive norms* have the form of "Do ...".

Coleman affirms that "proscriptive norms provide negative feedback in the system, damping out the focal actions; prescriptive norms provide positive feedback, expanding the focal action".

The author also claims that for any norm, “there is a certain class of actors whose actions or potential actions are the focal actions”.

Now that we have grounded some useful terms, we will proceed to analyze the different definitions of norms found in the literature.

DEFINING NORMS

As stated by Boella, van der Torre, and Verhagen (2008): “A normative multiagent system is a multiagent system organized by means of mechanisms to represent, communicate, distribute, detect, create, modify, and enforce norms, and mechanisms to deliberate about norms and detect norm violation and fulfilment”.

However, in the same work, authors make a differentiation depending on the point of view:

1. The *legalistic view* of normative multiagent systems is a top-down view which considers the normative system as a regulatory instrument to regulate emerging behavior of open systems without enforcing the desired behavior. Agents are often motivated by sanctions to stick to norms, rather than by their sharing of the norms. Even if agents are allowed some freedom to create norms, this freedom is mostly restricted to the possibility for agents to create contracts to regulate the interaction among them.
2. The *interactionist view* on normative multiagent systems represents a bottom-up view. In this autonomous, individually oriented view, norms can be seen as regularities of behavior which emerge without any enforcement system because agents conform to them either because their goals happen to coincide, or because they feel themselves as part of the group or because they share the values of other agents. Sanctions, or formal measures towards norm violating agents carried out by agents whose task it is to sanction norm violations, are not always necessary because social blame and spontaneous exclusion of non-conforming agents are often adequate to incentivize conformity to norms.

We have observed a lot of work in the literature that covers the legalistic point of view of a multi-agent system. We can also relate this legalistic point of view with systems where norms are predefined by an authority on the system, and it is done before runtime (Shoham & Tennenholtz, 1995; Boella, 2003; García-Camino, Noriega & Rodríguez-Aguilar, 2005).

We, however, are more interested in how norms emerge in a multiagent society. Hence, in this work we will focus on the interactionist point of view of multiagent systems. We cover norms that emerge in a decentralized process because of the interests and goals of the members of a society. Nevertheless the term social norm has been used by different areas of research. We can find several definitions of norms in the literature of economics, social sciences and multiagent systems.

The philosopher Hart (1961) defined *norms* as “a prescribed guide for conduct or action which is generally complied with by the members of a society”.

This definition, although very intuitive, is not complete. This definition affirms that norms are only the prescribed actions that members of a society should follow. However, we can find several examples of everyday-life norms that are proscriptive norms, e.g. don't smoke, don't drive on the wrong side of the road, don't wear inappropiate clothes to work, etc. Therefore, we need a definition of norm that cover both prescriptive and proscriptive norms.

One of the simplest definitions of norm is that used by Shoham and Tennenholtz (1997), who uses the term *social law* and defines it as "a restriction on the set of actions available to the agents. A game *g* and a social law *sl* induce a sub-game g_{sl} of *g* that is the restriction of *g* to actions that are not prohibited by *sl*."

In the same work, the authors also define *social convention*: "A social law that restricts the agents' behavior to one particular strategy is called a (social) convention".

Other relevant authors in multiagent and norm literature has also used this definition like Delgado, Pujol & Sangüesa (2003).

However this definition still lacks an important part of what norms are, and that is the existence of a sanction when not conforming to the norms.

Elster (1989) affirms that "norm-guided behavior is supported by the threat of social sanctions that make it rational to obey the norms". Axelrod (1986) uses a similar definition of norm: "A norm exists in a given social setting to the extent that individuals usually act in a certain way and are often punished when seen not to be acting in this way."

As the author claims, this definition "makes the existence of a norm a matter of degree, rather than all or nothing proposition, which allows one to speak about the growth or decay of a norm." Therefore a certain action will become a norm depending on "how often that action is taken, and how often one is punished for not taking it".

All the definitions presented so far are taken from the sociology literature. We can also find some definitions of social norms on the multiagent literature.

A definition, related to that of Axelrod presented previously which also deals with sanctions, is used by Boella and Lesmo (2002) (their definition is also borrowed from the sociologist Goffman) where they define norms as "a kind of guide to action that is supported by social sanctions". Moreover, they also specify that sanctions are defined "as a reaction of others to the behavior of an individual or a group, a reaction having the goal to enforce the respect of a given norm."

Coen (2000) uses the term social law instead but refers to the same kind of norms that we are interested. He affirms that "a social law is explicitly designed to prevent conflict and deadlock among the agents; however, for it to be deemed useful, it should simultaneously allow each agent to achieve its individual set of goals. [...] It must be sufficiently strict to prevent conflict or deadlock, and simultaneously, it must be sufficiently liberal to allow the agents to efficiently achieve their goals.". The definition used by this author is the following: "social laws are guidelines that specify a class of valid algorithms (or strategies) for solving problems from a particular domain by partitioning the set of possible algorithms into 'law-abiding' and 'criminal' sets."

Boella and Lesmo (2002) affirm that norms need to be represented “as a combination of beliefs and goals of the agent subjected to the norm, and of the agent who has to enforce the respect of the norm: in particular, the goal of avoiding sanctions, the goal of not violating the norm and the belief that there is another agent who has the goal of sanctioning violations”.

We observe that in the multiagent literature terms like social norms, social laws, and social conventions are used interchangeably. The main objective of this work is to clarify the differences and characteristics between disparate norms types.

After the review of the literature, we will unify all the definitions in the following: “A *social norm* is a restriction on the set of actions available to the agents, commonly shared by the members of the society and believed to be shared. The norm-followers act as enforcers by applying sanctions depending on the fulfilment of this norm.”

NORM OR CONVENTION?

After having presented the different definitions of norms, we analyze the different types of norms. In this section we will clasify the types of norms, which is determined by the type of problem that solve.

If we consider norms as regularities in the behavior of agents, we can observe two main types of norms:

1. Norms that naturally emerge, with no threat of punishment. These norms are called *conventions* or *conventional norms*. *Conventional norms* solve coordination problems, where there exist no conflict between the individual and the collective interests, as what is desired is that everyone behaves in the same way, without any major difference on which action agents are coordinated. Following Coleman’s theory, the selection of the focal action in such norms is arbitrary. One clear example of these kind of norms is the selection of which side of the road to drive on.
2. Norms that solve or ease collective action problems, where there is a conflict between the individual and the collective interests. These norms are called *essential norms*. Following the definition of focal actions, in these kind of norms the focal action is not chosen randomly because “the targets’ interests lie in the direction of action opposing observance of the norm, and the beneficiaries’ interests lie in the direction of action favoring observance of the norm” (Coleman, 1998).

Conventional Norms

Young (1993) presents the following definition of a *conventional norm*: “A convention is a pattern of behavior that is customary, expected, and self-enforcing. Everyone conforms, everyone expects others to conform, and everyone wants to conform given that everyone else conforms.”

This definition of conventional norm is perfectly compatible with that from Coleman (where he affirms that conventional norms solve the coordination problem of chosing equally beneficial focal points), and we will use this definition for *conventional norms*.

Essential Norms

On the other hand, the *essential norms* help address situations where the individuals are tempted not to contribute to the common good. These problems are commonly known in the literature as *collective action problems*. Heckathorn (1996) claims that any collective action system has two characteristics:

1. Non-excludable: Excluding anyone from consumption of the common good is impractical. For example, scabs benefit from higher wages won through strikes.
2. Jointness of supply: the degree to which the good costs the same to produce regardless of the number of people who consume it. A radio broadcast has very high jointness of supply because the costs of production have very little, if any, proportion to the number of people who consume the broadcast. Manufactured products, in contrast, typically have low jointness of supply because the cost of production increases with the number of consumers, though economies of scale may make the increase in cost less than directly proportional to the increase in consumers (Barros, 2007)

Referring to the first of these characteristics of collective action systems, Linares (2007) claims that in the case of the collective actions, conflicts appear when an individual is tempted not to contribute to the common pool, leading to an incomplete collective action. The fact that an individual behavior affects the welfare of the group is enough for the group to acquire the right to control individual behavior. Social norms are applied to control the individual behaviour in these kind of problems.

Prisoner's Dilemma and the Collective Action

It has been proven that in a game that follows the structure of a Prisoner's Dilemma [2] (like the one shown in Table 1) the individually rational strategy is to *Defect*, no matter what the other player decides. The equilibrium state of those decisions is suboptimal in the sense of Pareto, as there exists a different focal point (the mutually cooperative outcome) where the outcome when both agents are coordinated is preferred by both players. A social norm prescribing cooperation would help agents.

	B chooses	B chooses
	Cooperate	**Not Cooperate**
A chooses **Cooperate**	3 for A 3 for B	0 for A 5 for B

A chooses **Not Cooperate**	5 for A 0 for B	1 for A 1 for B

Table 1. The Prisoner's Dilemma

However, conflict between the individual's interest and the collective's interest can occur in other situation than that captured by the *Prisoner's Dilemma* game. Such conflicts can also have the form of the *Collective Action Game*, as shown in Table 2. Only the payoffs for the individuals are shown, and they are calculated considering that *V* is the value of the collective good, *c* is the cost for the individual to contribute, and *p* is the proportion of the total value that an individual can produce by itself.

	Collective choose **Cooperate**	Collective chooses **Not Cooperate**
Individual chooses **Cooperate**	*V - c*	*pV - c*
Individual chooses **Not Cooperate**	*V(1- p)*	*0*

Table 2. The Collective Action Game

We can observe that this *Collective Action Game* is equivalent to the *Prisoner's Dilemma* if the following conditions are fulfilled:

(1) $$V - c \le V(1 - p)$$

(2 $$pV - c \le 0$$

However, we can observe different types of collective action games depending on how the collective action is built. In some cases the collective action will be formed with a certain proportion of the population coordinated (and above that proportion it will not make a difference). There are other cases where there has to be at least a certain number of players cooperating so it is worth while to do so, and then, after that, the more agents the better.

Marwell and Oliver (1993) make a differentiation depending on how each unit of contributed resources affects the global collective good (therefore, how *p* changes with each contribution):

- **Linear Production Function**: Each unit of resource contributed to the common good produces the same outcome. Therefore, the slope in the production function is constant, i.e., *p* remains constant.
- **Decelerating Production Function**: Each unit of resources contributed to the common good produces less outcome when the donations increase, i.e., *p* decreases with each contribution.
- **Accelerating Production Function**: Each unit of resource contributed to the common good produces more outcome when the donations increase, i.e., *p* increases with each contribution.

Heckathorn (1996) affirms that the standard production function of a collective action game is an *S* shaped monotonically increasing curve. This *S* shaped curve contains at the same time the tree production functions defined by Marwell: the accelerating production function, the linear production function and the decelerating production function. Heckathorn also provides such a production function to calculate the level of the collective goods produced (*L*):

$$L = 1 - (D / N)^F \qquad (3)$$

where *N* is the number of actors in the group, *D* is the number of actors in the group who defect, and *F* is an exponent controlling the shape of the production function. Following this function, the level of of collective goods produced can vary from $L = 0$, or no production, to $L = 1$, indicating full production. (When intermediate numbers of actors contribute, i.e., $0 < D < N$, the link between the proportion contributing and the level of collective goods produced depends on the value of the exponent, *F*.)

- When $\boldsymbol{F = 1}$, the production function is linear. Contributions from any given proportion of the group produce that proportion of the collective good.
- When $\boldsymbol{F>1}$, the production function is decelerating. The first contributors are the most productive, while subsequent contributors face decreasing marginal returns. Therefore, in these situations, there are incentives for initial contributions, but 100% cooperation is rather difficult to achieve.
- When $\boldsymbol{F < 1}$, the production function is accelerating. These functions require near universal contribution to produce meaningful amounts of the collective good.

One example of an *S*-shape function is the following: imagine a society under a dictatorship and the majority members of this society are against this dictatorship but are also afraid of expressing their feelings against the repression. People starting and joining a demonstration against this dictatorship takes the form of a collective action game with an *S*-shape, with $F < 1$ initially. The first agents starting the demonstration will get a small reward (slowly accelerating) until a certain amount of people have joined the demonstration when rewards increase (accelerating), and then more and more people will join the demonstration in order not be ashamed by those already in the demonstration, although the effect of this people joining the demonstration will not have an important impact (decelerating), where *F > 1*.

Oliver and Marwell (1988) claim that the problem of the collective action is to find a subset of individuals with enough interest and resources to bootstrap the initial stages of the collective process in the area of slowly accelerating slope of the production function (where the initial contributions to the collective action have a low impact), in a way that after reaching the area where each new contribution is increasingly significant happen a snow-ball effect. However, we have to recall that the individuals in the critical mass (the leaders) are playing a totally different game than the rest of the population (the followers), because for the first ones the common good has a high value.

As it was analyzed previously, depending on how each unit of the contributed resource affects the collective good, the function will adquire a different shape. However, it can be analyzed from a different point of view: depending on the production function exponent (*F*) and on the relative value of the collective good (*V/c*), Heckathorn produced a figure (Figure 2) where confronts these two variables, obtaining different regions characterized by five different games. These games are: the prisoner's dilemma (represented in Table 1) , the chicken game (represented in Table 3) , the assurance game (represented in Table 4), the privileged game (represented in Table 5) and the altruist's dilemma game (represented in Table 6).

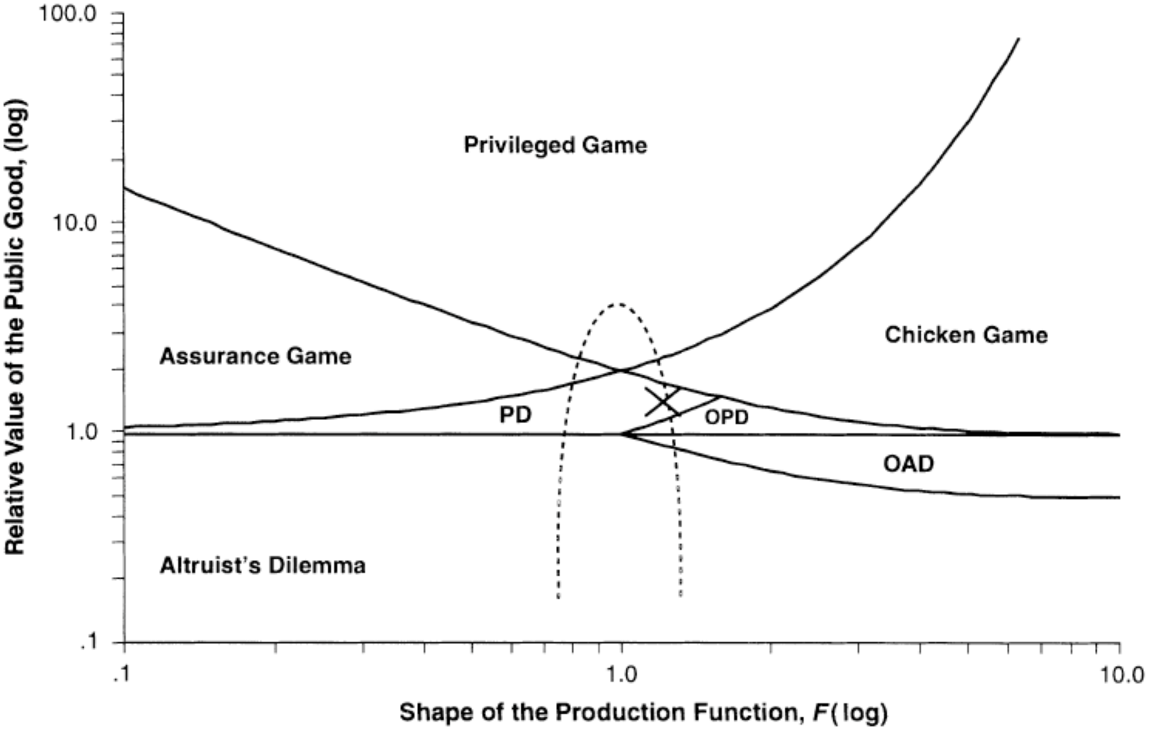

Figure 2. Heckathorn's Game-Space Diagram Showing the Family of Games Generated by the Relationship between the Shape of the Production Function (*F*) and the Relative Value of the Public Good (*V/c*)

Each of these games represents different scenarios where social norms are needed.

The chicken game represents games where the social norms regulate alternance: it is irrational that all the players cooperate, but only a subgroup of them, but which subgroup?

	B chooses	B chooses
	Cooperate	**Not Cooperate**
A chooses **Cooperate**	3 for A 3 for B	1 for A 5 for B
A chooses **Not Cooperate**	5 for A 1 for B	0 for A 0 for B

Table 3. The Chicken Game

The assurance game represent games where social norms regulate the participation of all the members, e.g, a demonstration is only worth it to be done if all the players demonstrate.

	B chooses	B chooses
	Cooperate	**Not Cooperate**
A chooses **Cooperate**	5 for A 5 for B	0 for A 3 for B
A chooses **Not Cooperate**	3 for A 0 for B	1 for A 1 for B

Table 4. The Assurance Game

The prisoner's dilemma represent scenarios known as the tragedy of the commons where social norms are needed for an ideal regulation.

Both the privileged and the altruist's dilemma game are trivial games where both full cooperation and full defection are the dominant strategies.

	B chooses **Cooperate**	B chooses **Not Cooperate**
A chooses **Cooperate**	5 for A 5 for B	1 for A 3 for B
A chooses **Not Cooperate**	3 for A 1 for B	0 for A 0 for B

Table 5. The Privileged Game

	B chooses **Cooperate**	B chooses **Not Cooperate**
A chooses **Cooperate**	1 for A 1 for B	0 for A 5 for B
A chooses **Not Cooperate**	5 for A 0 for B	3 for A 3 for B

Table 6. The Altruist's Dilemma Game

Collective Action Norms

So far we have analyzed the different strategies that agents can follow in order to fulfil (or to not) a norm. However, we need to analyze the different options agents have after this first decision.

Heckathorn (1996) differentiates two levels of collective actions (C.A.). *First Level* C.A. are contributions at the personal level to the collective good. E.g. go to a demonstration or help to build a canoe. *Second order* C.A. are those "such as selective incentives to reward first level cooperators or punish first order defectors".

Basically these two levels control, firstly, which actions agents take, and then, how do they observe the compliance of others' actions.

The author reason about six different strategies an agent can follow:

1. *Private Cooperation (CD)*: Agents contribute at the first level by contributing to the collective good but defect at the second level by forgoing any attempts to influence others.
2. *Full Defection (DD)*: Agents defect at both levels by refusing to contribute and permitting others to do as they wish.
3. *Full cooperation (CC)* involves contributing to collective goods production (first-level cooperation) and sanctioning those who fail to contribute (second-level cooperation)
4. *Hypocritical cooperation (DC)*: An actor defects at the first level but cooperates at the second level, failing to contribute to the collective good while acting to compel others to contribute.
5. *Compliant opposition (CO)* means cooperating at the first level but exercising oppositional control at the second level; the actor contributes to the collective good but defends the rights of others to refuse to contribute.
6. *Full opposition (DO)* means refusing to contribute and opposing norms that would compel compliance.

Consequently, we observe that sanctioning (either positively or negatively) first level actions plays a key role in the evolution of social norms. Therefore, in the next section we will analyze deeply the sanctioning mechanisms for norm defectors.

NORMS AND SANCTIONS

So far we have seen how norms are a useful coordination mechanism to solve a certain kind of problems in agent's societies. However, we understand that these norms also need reinforcement mechanisms that streghthen their fulfilment. Sanctioning is the most intuitive mechanism that will allow agents to reinforce the fulfilment of actions.

When talking about norms and sanctions a few questions arise: why are sanctions necessary? what is the objective of sanctioning a certain action? what types of sanctions exist? All these questions need to be answered when designing a norm-regulated multiagent system. In this section we will answer some of them in a general way.

First of all, it needs to be defined what a sanction is. In order to do that, we will borrow the definition of Coleman (1998): "If holding a norm is assumption of the right to partially control a focal action and recognition of other norm holders' similar right, then a sanction is the exercise of that right. A sanction may be negative, directed at inhibiting a focal action which is proscribed by a norm, or positive, directed at inducing a focal action which is prescribed by a norm."

From this definition we see that sanctions can be either positive or negative. However, we would like to add to this definition that sanctions have also a cost associated. This imposition cost have to be assumed by the agents applying the sanction. Combining the notion of the orientation of the sanction (positive or negative) and the cost associated to it (costly or zero-cost), we observe different types of sanctions.

The most straightforward positive sanction is the actual benefit obtained from abiding by the norm. For example, when driving in the side of the road specified by the convention, agents are reinforced positively because they do not crash with other agents.

Nonetheless the most common sanction observed in everyday life examples, as in the works on the literature, are negative sanctions. In a norm-abiding society, an agent should be punished when observed to be not following the norm. We can see how the fact of punishing improper behavior is an important mechanism for norms to be socially accepted. We can imagine several everyday situations where, if a reward system did not exist, a norm would not hold: admonish someone who jumps a queue, give a bad look when someone does not recycle, or ostracize the person that lights up a cigarette in a non-smoking environment.

Moreover, we would like to reckon special attention on an specific mechanism strongly related to social norms that can act either as a positive or a negative sanction (or both at the same time). This mechanism is *Reputation*. The cost assigned to reputation is the actual cost of communication and signalling, which in most of the cases is assumed to be zero or extremelly low (with respect to the other costly sanctioning mechanisms as it could be a fine system). The ability of other agents to observe if an agent has follow a norm affects to the own internal evaluations of that agent. Moreover, the ability of communication will allow agents to transmit this evaluations. Reputational effects are directly associated to the fulfilment of norms due to the ontological relationships between different norms: e.g. if a person has gained a good reputation in the community life of its neighborhood by recycling, it will directly obtain some privileges in other normative situations related to this community life. Similarly, reputation can affect negatively as well in the same way that positively.

Once that we have grounded the types of sanctions and how they affect the agents, we want to observe how these sanctions affect the emergence of norms. So far we have seen two main types of norms: the conventional norms and the essential norms.

The former norms do not need the help of external rewards in order for the norm be accepted by the society, as they are self-enforcing. The existence of external rewards can still accelerate the process of norm emergence.

For essential norms, however, external rewards are needed in order for the norm to be socially accepted.

These hypotheses are also shared by Young (2008), who claims that there are three mechanisms by which norms emerge:

- Pure Coordination: These are "social" phenomena, because they are held in place by shared expectations about the appropriate solution to a given coordination problem, but there is no need for social enforcement.
- Threat of social disapproval or punishment for norm violations.

- Internalization of norms of proper conduct.

The first mechanism is the one used in the conventional norms. The second mechanism is used in both conventional and essential norms and it is the one we are more interested in. Finally, the third mechanism is treated by Coleman (1998), who claims that "a norm may be embedded in a social system in a more fundamental way: the norm may be internal to the individual carrying out the action, with sanctions applied by that individual to his own actions. In such a case a norm is said to be internalized. An individual feels internally generated rewards for performing actions that are proper according to an internalized norm or feels internally generated punishments for performing actions that are improper according to an internalized norm."

Although we have analyzed the mechanisms by which norms emerge, we also need to understand the reasons why agents would decide to punish another agent. We need to analyze the sociology literature to understand these reasons. Banks (2009) help us analyze the reasons why punishment should be applied:

- Punishment will stop offenders from committing further crimes.
- Punishment tells the victim that society disapproves of the harm that he or she has suffered.
- Punishment discourages others from committing the same infractions.
- Punishment protects society from dangerous and dishonest people.
- Punishment allows an offender to make amends for the harm he or she has caused.
- Punishment ensures that people understand that laws are to be obeyed.

Once that we have clear what is the objective of punishment, we need to understand what are the characteristics of this punishment. Bean (1981) affirms that punishment consists of five elements:

1. It must involve an unpleasantness to the victim.
2. It must be for an offense, actual or supposed.
3. It must be the work of personal agencies; in other words, it must not be the natural consequence of an action.
4. It must be imposed by an authority or an institution against whose rules the offense has been committed. If this is not the case, then the act is not one of punishment but is simply a hostile act. Similarly, direct action by a person who has no special authority is not properly called punishment, and is more likely to be revenge or an act of hostility.

Furthermore, and now that it is clear the necessity of punishment mechanisms for norms to emerge, we need a classification of the possible punishments. Posner and Rasmusen (1999) claim that there exist the following types of sanctions:

1. Automatic Sanctions: Those that an agent receive for not being coordinated with the others.
2. Guilt: The violator feels bad about his violation as a result of his education and upbringing, quite apart from external consequences. Probably most people in our society, though certainly not all, would feel at least somewhat guilty about stealing even if they believed they were certain not to be caught.
3. Shame: The violator feels that his action has lowered himself either in his own eyes or in the eyes of other people. In its most common form, shame arises when other people

find out about the violation and think badly of the violator. The violator may also feel ashamed, however, even if others do not discover the violation. He can imagine what they would think if they did discover it, a moral sentiment which can operate even if he knows they will never discover it. Also, he may feel lowered in his own eyes, a “multiple self” situation in which the individual is both the actor and the observer of his actions.

4. Informational sanctions. The violator’s action conveys information about himself that he would rather others not know. A student wears casual clothing to a job interview, unintentionally signaling that he doesn’t really care about getting the job.
5. Bilateral costly sanctions. The violator is punished by the actions of and at the expense of just one other person, whose identity is specified by the norm. The expense to that person could be the effort needed to cause the violator disutility, or the utility that the person imposing the punishment loses by punishing him. Examples of what we are calling bilateral costly sanctions are where an adulterer is shot by a jealous husband and where the husband divorces his wife after discovering her adultery.
6. Multilateral costly sanctions. The violator is punished by the actions and at the expense of many other people. A divorced man finds that he is no longer invited to dinner in the community.

CONCLUSION

The aim of this article has been to better understand and characterize the different types of social norms. Accordingly, we have discussed several approaches to social norms from the most relevant research areas including sociology, economics and multiagent systems.

We have defined some terms that have helped us better understand social norms and the related research. We have also provided a deeper analysis on the definitions that has been presented introducing our own definition of social norm.

We next used a game-theoretical characterization to derive a classification of the different types of norms, which will also help to accurately categorize social norms in the future. We have seen how conventional norms (norms that help selection of a focal action from the different possible focal actions) are different from essential norms (norms that promote the adoption of the optimal strategy). In addition, a thorough study was undertaken to distinguish the different types of essential norms by analyzing the search space of the collective action game.

Finally, we presented an analysis of the sanctioning mechanisms and how these can affect the emergence of norms. We have presented a setof characteristics common to sanctions that have to be considered when designing a normative multiagent system. However, there are a number of questions that still need to be answered: how can agents detect an improper behaviour? Who is in charge of applying the sanction and incur the associated cost? How can we ensure that agents will efficiently learn new norms?

In summary, in this article we have seen how research in normative multiagent systems has gained momentum in the last few years. Yet, there are several important unanswered research questions that needs to be adequately addressed.

Acknowledgments

This work was supported by the Spanish Education and Science Ministry [AEI project TIN2006-15662-C02-01, AT project CONSOLIDER CSD2007-0022, INGENIO 2010]; Proyecto Intramural de Frontera MacNorms [PIFCOO-08-00017] and the Generalitat de Catalunya [2009-SGR-1434]. Daniel Villatoro is supported by a CSIC predoctoral fellowship under JAE program. Sandip Sen is partially supported in part by a DOD-ARO Grant #W911NF-05-1-0285.

Biographies

Daniel Villatoro is a PhD student at the Artificial Intelligence Research Institute (IIIA) of the Spanish National Research Council (CSIC) in Barcelona, Spain. His research interests include decentralized control mechanisms and the topology of virtual societies, particularly in applying social norms to distributed Systems. He received his MSc in Computer Science from the University of Granada. He has won several scholarships that have allowed him to visit different universities and research departments such as the University of Ottawa, the Politecnico di Milano, the Santa Fe Institute or the University of Tulsa.

Sandip Sen is a Professor of Computer Science in the University of Tulsa with primary research interests in multiagent systems, machine learning, and genetic algorithms. He completed his PhD in the area of intelligent, distributed scheduling from the University of Michigan in December, 1993. He has authored approximately 200 papers in workshops, conferences, and journals in several areas of artificial intelligence. In 1997 he received the prestigious CAREER award given to outstanding young faculty by the National Science Foundation. He has served on the program committees of most major national and international conferences in the field of intelligent agents including AAAI, IJCAI, ICMAS, AA, AAMAS, ICGA, etc. He was the co-chair of the Program Committee of the 5th International Conference on Autonomous Agents held in Montreal Canada in 2001. He has chaired multiple workshops and symposia on agent learning and reasoning. He has presented several tutorials on multiagent systems in association with the leading international conferences on autonomous agents and multiagent systems.

Jordi Sabater-Mir is tenured scientist at the Artificial Intelligence Research Institute (IIIA) of the Spanish National Research Council (CSIC), Barcelona, Spain. He holds a doctorate in Artificial Intelligence and has been a postdoctoral Marie Curie fellow at the Institute of Cognitive Sciences and Technologies (ISTC-CNR) in Rome, Italy. His current research is focused on computational models of trust and reputation, agent based social simulation (normative systems), development of cognitive agents and electronic institutions. He has published more than 60 papers in journals and international conferences and has participated in several European and national research projects (some of them as main researcher). He has been

PC member of the main conferences and workshops in the area of MAS and has organized several workshops in the area of computational trust and reputation systems.

REFERENCES

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

Banks, C. (2009). *Criminal Justice Ethics. Theory and Practice*. Thousand Oaks, CA: SAGE Publications.

Barros, B. (2007). *Group Size, Heterogeneity, and Prosocial Behavior: Designing Legal Structures to Facilitate Cooperation in a Diverse Society*. SSRN eLibrary.

Bean, P. (1981). *Punishment: A Philosophical and Criminological Inquiry*. Oxford, UK: Martin Robertson.

Boella, G. (2003). Norm governed multiagent systems: The delegation of control to autonomous agents. In *Proceedings of the IEEE/WIC Intelligent Agent Technology Conference* (pp. 329-335). Washington, DC: IEEE Computer Society.

Boella, G., & Lesmo, L. (2002). A game theoretic approach to norms and agents. *Cognitive Science Quarterly,* 492-512.

Boella, G., van der Torre, L., & Verhagen, H. (2008). Introduction to the special issue on normative multiagent systems. *Autonomous Agents and Multi-Agent Systems, 17*(1), 1-10.

Coen, M. H. (2000). Non-deterministic social laws. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (pp. 15-21). Cambridge, MA: AAAI Press/MIT Press.

Coleman, J. (1998). *Foundations of social theory*. Cambridge, MA: Belknap Press.

Delgado, J., Pujol, J. M., & Sangüesa, R. (2003). *Emergence of coordination in scale-free networks*. Web Intelligence and Agent Systems 1(2), 131–138.

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

García-Camino, A., Noriega, P., & Rodriguez-Aguilar, J. A. (2005). Implementing norms in electronic institutions. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '05)* (pp. 667-673). New York: ACM.

Hart, H. L. (1961). *The concept of Law*. Oxford, UK: Oxford University Press

Heckathorn, D. D. (1996). The dynamics and dilemmas of collective action. *American Sociological Review, 61*(2), 250-277.

Linares, F. (2007). The problem of the emergence of social norms in collective action. an analytical approach. *Revista Internacional de Sociología, 65*(46), 131-160.

Marwell, G., & Oliver, P. (1993). *The Critical Mass in Collective Action: A Micro-Social Theory*. Cambridge, UK: Cambridge University Press

Oliver, P., & Marwell G. (1988). The paradox of group size in collective action: A theory of the critical mass. *American Sociological Review, 53*, 1-8.

Posner, R., & Rasmusen, E. (1999). Creating and enforcing norms, with special reference to sanctions. *Law and Economics, 19.3*(1999), 369-382.

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

Shoham, Y., & Tennenholtz, M. (1997). On the emergence of social conventions: Modeling, analysis, and simulations. *Artificial Intelligence, 94*, 139-166.

Young, H. P. (1993). The evolution of conventions. *Econometrica, 61*(1), 57-84.

Young H. P. (2008). Social norms. In *The New Palgrave Dictionary of Economics*. New York: Palgrave Macmillan.

[1] Sanctioning is an important mechanism by which social norms emerge and are reinforced in the society.

[2] The Prisoner's Dilemma states: Two suspects are arrested by the police. The police have insufficient evidence for a conviction, and, having separated both prisoners, visit each of them to offer the same deal. If one testifies (defects from the other) for the prosecution against the other and the other remains silent (cooperates with the other), the betrayer goes free and the silent accomplice receives the full 5-year sentence. If both remain silent, both prisoners are sentenced to only six months in jail for a minor charge. If each betrays the other, each receives a five-year sentence. Each prisoner must choose to betray the other or to remain silent. Each one is assured

that the other would not know about the betrayal before the end of the investigation. How should the prisoners act?