An Ontology of Social Control Tools

Philippe Pasquier
DAMAS Laboratory
Laval University
G1K 7P4, Québec, Canada
pasquier@iad.ift.ulaval.ca

Roberto A. Flores
Christopher Newport
University
Newport News, VA 23606
flores@pcs.cnu.edu

Brahim Chaib-draa
DAMAS Laboratory
Laval University
G1K 7P4, Québec, Canada
chaib@iad.ift.ulaval.ca

ABSTRACT

In multi-agent systems, social commitments are increasingly used to capture roles, social norms, the semantics of agent communication as well as other inter-agent dependencies. Those systems rest on the assumption that agents respect their commitments. In this paper, we present an ontology of sanctions and punishment philosophies which are required ingredients of any social control mechanism susceptible to fosters agents' compliance with the commitments they create.

1. INTRODUCTION AND MOTIVATIONS

A multi-agent system (MAS) is considered an *open* MAS if the following properties hold [5]: (1) agents' behavior and interactions cannot be predicted in advance; (2) agents' internal architecture is not publicly known and (3) agents do not necessarily have common goals, desires or intentions.

The first of those properties implies that the execution of open MAS is *non-deterministic*. Open societies are usually subject to unanticipated outcomes in their interactions. The second property implies that an open MAS can have members with different internal architectures; therefore, they can be *heterogeneous*. The third property implies that the members of an open society may be non-benevolent, non-cooperative or even insincere. In addition, the agents may fail to, or choose not to, conform to some of the normative aspects of the MAS in order to achieve their individual goals. In that context, providing the means and tools for the achievement of a chosen/emergent social order in such system is a challenging issue.

Social commitment has been introduced as a public and social primitive that allows to capture all social and normative dimensions of MAS, including: social norms, roles and the semantics of communication as well as other interagent dependencies. One of the key feature of those public and social primitives is their flexibility, which allow to take into account the dynamic of the environment. Most approaches assume that the agents will respect their social commitments (thus applying regimentation). This assumption is unrealistic since unintended violation is likely to occur and unilateral cancellation as well as commitment modification are desirable.

In a perspective of knowledge transfer, this short paper presents an ontology of social control mechanisms that can be used to support the *enforcement of flexible social commitments* in open systems.

2. ONTOLOGY OF SANCTIONS AND SOCIAL CONTROL MECHANISMS

Introduced in sociology as early as the end of the 19th century, the concept of social control originally denoted the capacity of a group or society to regulate itself and to secure coherency and unity in social life [7]. Social control, in this sense, relates to how social action is coordinated toward a chosen or an emergent social order. Often seen as all-encompassing, practically representing any phenomenon leading to conformity or as a broad representation of regulated mechanisms placed upon society's members, social control can be viewed as the glue holding society together [4].

Modern theories of social control focus on the strategies and techniques that help to regulate agent behavior, and lead to conformity and compliance with the rules of society (at both the macro and the micro levels). In the remainder of this section, we detail the main elements used in the enforcement of social commitments: (1) sanctions, which are considered in their general sense of incentives (the next section presents an ontology of sanctions along their different dimensions), and (2) philosophies of punishment (Section 2.2), which result in punishment strategies determining the type of sanction (and its magnitude) to be applied, and explains how sanctions are assigned to social commitments.

2.1 Sanctions

In this paper, we only consider *individual sanctions* and, for simplicity, leave aside other types of sanctions, such as *collective sanctions* [6] (which may be associated to teams, roles or groups of agents). In the next subsections, we go through the three main dimensions of sanctions: direction, type and style.

Sanction Directions

Sanctions have a specific direction. It is usually useful to consider both:

- *positive sanctions*: positive sanctions are rewards that encourage a continuation of desired behavior. For ex-

AAMAS'06 May 8–12 2006, Hakodate, Hokkaido, Japan.

ample, it is common in open systems that agents accept committing to a task only if the associated reward is worth pursuing.

- *negative sanctions*: on the other hand, negative sanctions are used to discourage norm violating behavior. For example, agents that cannot fulfill their commitments are expected to be punished.

In brief, positive sanctions are incentives to pursue a particular behavior while negative sanctions are incentives against its violation. For the sake of simplicity, in the rest of the paper we use the words sanction to denote negative sanctions, and reward to denote positive rewards

Sanction Types

The first sanction type is *automatic sanctions*, which arises when the violators action carries its own penalty (e.g., because it is not being coordinated with the actions of others). For example, someone who drives on the wrong side of the road has a higher than normal probability of crashing into another car. We will not consider these unintended (since no one decides that they should apply) sanctions in the scope of this paper.

Within the vast literature addressing this topic from various perspectives including economics, criminology, sociology, social psychology, AI and MAS, we encounter three broad types of non-automatic sanctions: (1) material sanctions, (2) social sanctions, and (3) psychological sanctions.

Material sanctions include physical sanctions like physical retaliation or repairing actions, as well as financial sanctions like fees. Material sanctions can be applied immediately at the time of occurrence or be delayed through time.

There are *social sanctions* as well. Trust, credibility and reputation are social values that could be affected by social sanctions. As pointed out in [9], social sanctions are usually the effects of some implicit informational disclosure where the violator's action conveys information about himself that he would rather not have others know. For example, that an agent cancels or modifies a big number of its commitments without any explicit reason might be taken into account by the other agents when evaluating his reputation and the trust they put in him.

Psychological sanction types, which may be more useful in believable agents [1] and which have been used in advanced mono-agent design in mixed communities, can be important as well. Examples of psychological sanctions are guilt (where the violator feels bad about his violation as a result of his knowledge of social norms, quite apart from external consequences), and shame (where the violator feels that his action has lowered himself either in his own eyes or in the eyes of other agents).

The time horizon of sanctions indicates whether the effects of sanctions are long-lasting or short-lived. This concept is important since some sanction types may extend through time (e.g., trust, reputation, credibility) while others may not (e.g., immediate material sanctions). Subtle and complex phenomena, like forgiveness, can require taking into account this time issue.

Sanction Styles

For the specific formal needs of MAS, we distinguish two sanctions styles: implicit and explicit. *Implicit sanctions* are "autonomously" and unilaterally decided by agents. The major difficulty associated with implicit sanctions is that they are not publicly known and agents have to discover whether or not they have been sanctioned (for example, by noticing that others do not communicate with them anymore). On the contrary, *explicit sanctions* are publicly known (at least among the interacting agents).

Another useful distinction can be made between *a priori* decided sanctions and *a posteriori* decided sanctions. In particular, a posteriori decided sanctions do not allow agents to reason about the pros and cons of respecting their commitments. That the punished agent can disagree with the sanction assigned a posteriori may lead to litigation. The complexity associated with litigation makes a posteriori sanctions less desirable than a priori sanctions in open MAS.

In the remainder of the paper, we will consider only a priori defined explicit sanctions. Among a priori known explicit sanctions, we can distinguish *static*, a priori known, explicit, sanction systems provided to all agents at design time, and *dynamically* decided, a priori, explicit sanctions, which are negotiated by the agents through their communications.

2.2 Punishment Policies

Social control mechanisms to enforce social commitments should be designed according to a philosophy of punishment. By punishment, we mean the imposition of sanctions to satisfy open system designers' desire for retribution against wrongdoers. According to social control theorists, there are five different philosophies of punishment from which all *punishment policies* can be derived [10]:

- *Deterrence*: issued from the classical school of criminology, and supported by philosophers like Beccaria [2] and Bentham [3], deterrence is a utilitarian principle stating that the aim of sanctions is to prevent future violation. For deterrence to be effective, punishment must be swift, certain and severe. Applied to the enforcement of social commitment in MAS, it means that commitments should be associated with heavy and explicit sanctions. This extreme position, i.e. using severe sanctions with a high prohibitive effect, tends to transform social commitments into strict obligations, losing part of the flexibility objective desired for commitments.

- *Retribution*: retribution considers that the violation should be repaired by a penalty as severe as the wrongful act.

- *Incapacitation*: incapacitation considers the impairment or restriction of the agent's ability to commit the violation again. In human societies: exclusion, firing, and imprisonment, are the most common methods of incapacitation. Applied to the enforcement of social commitment in MAS, this punishment philosophy may result, for example, in the exclusion of wrongdoers from the system (at least for a certain time). As in human societies, it could have a preventive effect toward others violations, but it is not sure that it could be practical for artificial agents. Although, it is a consideration that is system-dependant, incapacitation usually results in the loss of opportunity during the time of the punishment, a circumstance that could be considered as a material sanction. For example, in e-business systems, losing activity time usually has material consequences. Since it is reducible to some material explicit sanction, we will consider incapacitation as a special case of the retribution philosophy in the rest of the paper.

- *Rehabilitation*: rehabilitation is the process of correcting from erroneous behavior and returning to a rightful course. This philosophy seems more difficult to implement in MAS since it would require policies to notify the entity responsible for wrongdoers that its decision-making mechanism must be corrected. Moreover, this philosophy would suppose pro-social learning capabilities that cannot be imposed in an heterogeneous open system.

- *Restoration*: restoration attempts to make the victim and community "whole again" through rituals. This punishment mainly occurs in less industrialized countries and will not be treated here since it requires social and cultural dimensions that are not yet considered in MAS.

Since punishment philosophies like incapacitation, rehabilitation and restoration focus on the choice of sanctions types and styles rather than on the choice of sanctions strength, we will restrict our analysis to the two remaining philosophies: deterrence and retribution. While both of them seem adequate for open MAS, we will argue that retribution is a practically better choice for open MAS.

Indeed, the last decades of work in economics and law provide two basic reasons why it is best for sanctions to equal harm[1]. Here, we reformulate these arguments toward retribution punishment policies using MAS terminology.

The first argument concerns the *level of precautions* taken by parties, where the term "precautions" is to be interpreted generally. If sanctions are less than harm, precautions will tend to be inadequate and agents will tend to not respect adopted social commitments when it is to their advantage to do so. Symmetrically, if sanctions exceed harm, precautions will be excessive and may preclude agents committing to wanted commitments (this is the case with the deterrence punishment philosophy). For example, even if sincerely wanting to, an agent will not commit to a course of action if the sanctions attached to violation (which may occur unintentionally) are prohibitive. However, it has been shown that if sanctions equal harm, agents will have socially correct incentives to take precautions [8].

The second reason why it is desirable for sanctions to equal harm involves the agents' *level of activity*, that is, the extent to which agents participate in risky activities. An agent's level of activity affects the magnitude of expected total harm, independently of the precautions taken when engaging in an activity. For example, the more commitments an agent takes (its level of activity), the greater the possible number of accidents (violations) will occur, independently of the safety features of the agent (which affect the expected harm per commitment) [8].

It is worth noticing that concluding that damages should equal harm would require making two assumptions. The first assumption is that agents are *risk neutral*. If injurers are risk averse then the optimal level of sanctions tends to be lower than harm, because it reduces the imposition of risk on injurers and because sanctions do not need to be as high to induce injurers to behave appropriately. The second assumption is that of *strict liability*, which stipulates that injurers are definitely found liable and that injurers cannot escape the corresponding sanctions.

3. CONCLUSION

In this paper, we have raised the problem of the enforcement of flexible social commitment in open systems which has been neglected in the past. We have introduced and discussed tools for treating this problem (Section 2), namely sanctions (Section 2.1) and punishment philosophies (Section 2.2). In doing so, we have shown that supporting the enforcement of social commitments may be made through a priori defined explicit sanctions (statically specified or dynamically negotiated) under a retribution punishment philosophy. In order to gain flexibility, the regimentation usually associated with social commitments has been moved to the sanction system itself. Indeed sanctions must be respected to ensure that this solution does not lead to the meta-problem of the enforcement of sanctions.

Notice that the design of (domain dependent) punishment policies, linking objective (or subjective) actions values to material sanctions, is an open research issue that we wish to inquire in future work.

4. REFERENCES

[1] J. Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, July 1994.

[2] C. Beccaria. *On Crimes and Punishments*. New Jersey: Prentice Hall, 1963.

[3] J. Bentham. *An Introduction to the Principles of Morals and Legislation*. London: The Athlone Press, 1970.

[4] M. Hechter and K. D. Opp. Introduction. In M. Hechter and K. D. Opp, editors, *Social Norms*, pages xi–xx. Russell Sage Foundation, 2001.

[5] C. Hewitt. Open information systems semantics for distributed artificial intelligence. *Artificial Intelligence*, 47:76–106, 1991.

[6] D. J. Levinson. Collective sanctions. Public law research paper no. 57, New York University, School of Law, Center for Law and Business Research, 2003.

[7] D. Martindale. *Social Control for the 1980s: A Handbook for Order in a Democratic Society*, chapter The theory of social control, pages 46–58. Westport, CT: Greenwood Press, 1978.

[8] M.A. Polinsky and S. Shavel. *The New Palgrave Dictionary of Economics and The Law*, volume 3, chapter Punitive Damages, pages 192–198. London: Macmillan Reference Limited, 1998.

[9] R. A. Posner and E. B. Rasmusen. Creating and enforcing norms, with special reference to sanctions. *International Review of Law and Economics*, 19(3):369–382, 1999.

[10] G. B. Vold, T. J. Bernard, and J. B. Snipes. *Theoretical Criminology*. Oxford University Press, 5^{th} edition, 2002.

[1] Here, harm is the violation of a particular social commitment and is at least equal to the effort that is needed to fulfill the commitment.