

# Enforceable Social Laws

Guido Boella  
Dip. di Informatica  
Università di Torino  
Italy  
guido@di.unito.it

Leendert van der Torre  
CWI Amsterdam  
and Delft Univ. of Technology  
The Netherlands  
torre@cwi.nl

## ABSTRACT

In this paper we study the enforcement of social laws in artificial social systems using a control system. We define the enforceable social law problem as an extension of Tennenholtz' stable social law problem. We distinguish the choice of social laws from the choice of control systems, where the latter leads to new computational problems. We consider also properties of sanction based control systems, and monitoring when there is no full observability.

## Categories and Subject Descriptors

I.2.11 [Distributed Artificial Intelligence]: Multi-agent systems

## General Terms

MAS theory

## Keywords

Artificial social systems, normative systems, normative multi-agent systems

## 1. INTRODUCTION

The basic idea of the artificial social systems approach is to add a mechanism, called a social law, that will minimize the need for both centralized control and on-line resolution of conflicts. A social law is defined as a set of *restrictions* on the agents' activities which allow them enough freedom on the one hand, but at the same time constrain them so that they will not interfere with each other [16].

Social laws may be seen as a kind of social norms, and artificial social systems may therefore be seen as normative multi-agent systems, that is, "sets of agents (human or artificial) whose interactions can fruitfully be regarded as norm-governed; the norms prescribe how the agents ideally should and should not behave" [12]. Moreover, norms have also been studied in deontic logic, a branch of philosophical logic that studies logical relations among obligations,

permissions and prohibitions. Deontic logic has been developed to model legal and moral systems, though it has also been applied to problems in computer science that involve soft constraints, i.e., constraints that can be violated [17].

Shoham and Tennenholtz [14] introduce off-line design of *useful* social laws for artificial agent societies, and Fitoussi and Tennenholtz [10] distinguish two criteria to choose social laws called *minimal* and *simple* social laws. Shoham and Tennenholtz [15] study the emergence of *rational* social laws in repeated games instead of their off-line design. However, these approaches can be criticized, because they do not incorporate the possibility that social laws can be violated [1].

This criticism has been addressed by Briggs and Cook [8], who introduce so-called *flexible* social laws that can be violated if an agent cannot obey the law. However, it is assumed that agents obey the laws whenever possible. Moreover, Tennenholtz [16] introduces *stable* social laws as a kind of qualitative equilibrium, in the sense that agents can deviate from the law, but they do not want to do so when the other agents follow it. This approach bridges social laws with conflict resolution. Brafman and Tennenholtz [7] study efficient learning equilibria in repeated games. However, such flexible and stable laws can be criticized too, because they assume that social laws are followed by agents voluntarily, whereas normative multi-agent systems are based on a control system. For example, in legal systems norms can be enforced and violations can be sanctioned.

In this paper we therefore address the following three questions:

1. How can we extend artificial social systems with a control system such that enforceable social laws can be defined? We model a normative system as an agent as proposed by Boella and Lesmo [3].
2. How to model the monitoring of violations? We extend the model to deal with the possibility that violations are not observed, or that sanctions cannot be enforced.
3. Which control system to choose, if there are several alternatives? For example, in principle very high sanctions can be added for any violation, but from the social sciences it is known that high sanctions for minor offences have drawbacks [2].

The layout of this paper is as follows. In Section 2 we discuss Tennenholtz notion of stable social laws in artificial social systems, in Section 3 we discuss control systems in normative multi-agent systems, and in Section 4 and 5 we introduce enforceable social laws. In Section 6 we discuss the choice of the control system.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

AAMAS'05, July 25-29, 2005, Utrecht, Netherlands.

Copyright 2005 ACM 1-59593-094-9/05/0007 ...\$5.00.

## 2. ARTIFICIAL SOCIAL SYSTEMS

The problem studied in artificial social systems is the design, emergence or more generally the creation of social laws. Shoham and Tennenholtz [14] introduce social laws in a setting without utilities. Since we need utilities to explain why agents follow the law, we use the kind of social laws introduced by Shoham and Tennenholtz in [15]. They define *rational* social laws as social laws that improve a social game variable. All definitions in this section are taken from Tennenholtz' presentation of artificial social systems for stable social laws [16].

### 2.1 Game

A game or multi-agent encounter is a set of agents with for each agent a set of strategies and a utility function defined on each possible combination of strategies. Tennenholtz only defines games for two agents to keep the presentation of artificial social systems as simple as possible, but he also observes in [16, footnote 4] that the extension to the multi-agent case is straightforward. For the same reason, in this paper we consider only games with two agents (extended with a normative system).

DEFINITION 1. A game (or a multi-agent encounter) is a tuple  $\langle N, S, T, U_1, U_2 \rangle$ , where  $N = \{1, 2\}$  is a set of agents,  $S$  and  $T$  are the sets of strategies available to agents 1 and 2 respectively, and  $U_1 : S \times T \rightarrow \mathbb{R}$  and  $U_2 : S \times T \rightarrow \mathbb{R}$  are utility functions for agents 1 and 2, respectively.

We use here as game variable the maximin value, following Tennenholtz [16]. This represents safety level decisions, in the sense that the agent optimizes its worst outcome assuming the other agents may follow any of their possible behaviors. See Tennenholtz' paper for a discussion why this is natural in many multi-agent systems, where a payoff corresponds to the achievement of a particular user's specification.

DEFINITION 2. Let  $S$  and  $T$  be the sets of strategies available agent 1 and 2, respectively, and let  $U_i$  be the utility function of agent  $i$ . Define  $U_1(s, T) = \min_{t \in T} U_1(s, t)$  for  $s \in S$ , and  $U_2(S, t) = \min_{s \in S} U_2(s, t)$  for  $t \in T$ . The maximin value for agent 1 (respectively 2) is defined by  $\max_{s \in S} U_1(s, T)$  (respectively  $\max_{t \in T} U_2(S, t)$ ). A strategy of agent  $i$  leading to the corresponding maximin value is called a maximin strategy for agent  $i$ .

### 2.2 Social laws

In the traditional so-called *useful* social law problem defined by Shoham and Tennenholtz [14], both the state with and without the social law are traditional games; the only distinction is that the sets of strategies of the agents in the former game are subsets of the ones in the latter game. A social law has therefore been characterized as a restriction of the strategies available to the agents. It is *useful* with respect to an efficiency parameter  $e$  if each agent can choose a strategy that guarantees it a payoff of at least  $e$ .

DEFINITION 3. Given a game  $g = \langle N, S, T, U_1, U_2 \rangle$  and an efficiency parameter  $e$ , we define a social law to be a restriction of  $S$  to  $\bar{S} \subseteq S$ , and of  $T$  to  $\bar{T} \subseteq T$ . The social law is useful if the following holds: there exists  $s \in \bar{S}$  such that  $U_1(s, \bar{T}) \geq e$ , and there exists  $t \in \bar{T}$  such that  $U_2(\bar{S}, t) \geq e$ . A (useful) convention is a (useful) social law where  $|\bar{S}| = |\bar{T}| = 1$ .

A social law is *quasi-stable* if an agent does not profit from violating the law, as long as the other agent conforms to the social law (i.e., selects strategies allowed by the law).

DEFINITION 4. Given a game  $g = \langle N, S, T, U_1, U_2 \rangle$ , and an efficiency parameter  $e$ , a quasi-stable social law is a useful social law (with respect to  $e$ ) which restricts  $S$  to  $\bar{S}$  and  $T$  to  $\bar{T}$ , and satisfies the following: there is no  $s' \in S - \bar{S}$  which satisfies  $U_1(s', \bar{T}) > \max_{s \in \bar{S}} U_1(s, \bar{T})$ , and there is no  $t' \in T - \bar{T}$  which satisfies  $U_2(\bar{S}, t') > \max_{t \in \bar{T}} U_2(\bar{S}, t)$ .

A quasi-stable social law is a *stable social law* if the payoff guaranteed to each of the agents is independent of the strategy (conforming to the law) it selects, as long as the other agent conforms to the social law. Tennenholtz motivates the use of stable social laws as follows.

“A quasi-stable social law will make a deviation from the social law irrational as long as the other agent obeys the law. However, the above definition of stability may not be satisfactory in our context. In a multi-agent encounter an agent has specific goals to obtain, and there is no reason to assume an agent will execute a strategy which yields to it a payoff guaranteed to it by another strategy, assuming the other agent obeys the law. [...] There is no reason to include in the set of allowed strategies a strategy which is (maximin) dominated by another strategy in the set.”

However, the existence of a quasi-stable social law does not imply the existence of a stable social law, which may be an argument to prefer for some applications the use of quasi-stable social laws.

DEFINITION 5. Given a game  $g = \langle N, S, T, U_1, U_2 \rangle$  and an efficiency parameter  $e$ , the quasi-stable social law that restricts the agents to  $\bar{S}$  and  $\bar{T}$  respectively is a stable social law if: for all  $s_1, s_2 \in \bar{S}$ , and  $t_1, t_2 \in \bar{T}$ , we have that  $U_1(s_1, t_1) = U_1(s_2, t_2)$  and  $U_2(s_1, t_1) = U_2(s_2, t_2)$ .

Note that there is no clear distinction between the state before the social law was introduced, and after it. In the traditional approach of Shoham and Tennenholtz [14], the creation of a social law leads to a new game, of which again the sets of strategies can be further reduced by a new social law. In the stable social law approach, the social law is implicit in the assumption of the agents that other agents follow it.

### 2.3 Computational problem

Finally Tennenholtz defines the computational problem he studies, called the stable social law problem or SSLP.

DEFINITION 6 (SSLP). Given a multi-agent encounter  $g$ , and an efficiency parameter  $e$ , find a Stable Social Law which guarantees to the agents a payoff which is greater than or equal to  $e$  if such a law exists, and otherwise announce that no such law exists.

Tennenholtz proves NP completeness for his problem SSLP, and he shows that the SSLP when one of the agents is logarithmically bounded is polynomial. He also gives a graph-theoretic representation of stable social laws.

### 3. NORMATIVE MULTI-AGENT SYSTEMS

In this section we discuss insights from normative multi-agent systems and deontic logic, which we use in the following sections in artificial social systems.

#### 3.1 Violability of norms

It is commonly understood in normative multi-agent systems and deontic logic that the violability of norms is of central importance: “Importantly, the norms allow for the possibility that actual behavior may at times deviate from the ideal, i.e., that violations of obligations, or of agents’ rights, may occur” [12]. There are many reasons why norms can be violated. There might be conflicts among norms and norms also cannot predict and successfully frame all possible circumstances. There might be some important event or fact to be handled, where no norm applies or some norm applies with bad results [9]. Shoham and Tennenholtz [14] acknowledge the relevance of this issue for artificial social systems.

“Harder assumptions to relax include the assumption of law-abiding citizenship; here we have assumed that all agents obey the social laws, but what happens if some don’t? How vulnerable is the society to rogue agents?” [14]

#### 3.2 Norm creation

Several models of norm creation have been proposed. For example, our model of the social delegation cycle [5] models the creation of social norms in the following three steps.

**Social goal generation.** From the individual desires of the agents we can derive joint (or social, or group) goals. For example, from the desire to be safe we can derive the joint goal that we want a safe society.

**Norm generation.** From a social goal, the norm generation process generates a norm for individual agents, together with a control system. For example, from the joint goal that we want a safe society, we can derive the norm not to kill, otherwise the killer is put in prison.

**Norm acceptance.** The norm with its control system must be accepted by the agents. For example, I accept the norm not to kill, which restricts my freedom to kill, because it achieves my desire to be safe if everyone follows the norm.

Note that the norm acceptance criterion represents a kind of normative equilibrium. It is an extension of the qualitative equilibrium condition encoded in stable social laws and discussed in Section 2, because an additional condition is that for each agent (or at least, for most agents) the social law is better than the original state.

#### 3.3 Normative system as agent

Boella and Lesmo [3] formalize the normative system as an agent to study games between ordinary agents and the normative system. The idea is borrowed from Goffman’s strategic interaction [11]. In a normative system, the “enforcement power is taken from mother nature and invested in a social office specialized for this purpose, namely a body of officials empowered to make final judgements and to institute payments” [11, p.115]. Such a game is unusual since

“the judges and their actions will not be fully fixed in the environment, many unnatural things are possible. [...] the payment for a player’s move ceases to be automatic but is decided on and made by the judges” [11, p.115]. “Strategic interaction” here means the, according to Goffman unavoidable, taking into consideration of the other agents’ actions.

“When an agent considers which course of action to follow, before he takes a decision, he depicts in his mind the consequences of his action for the other involved agents, their likely reaction, and the influence of this reaction on his own welfare” [11, p. 12].

Besides modelling the normative system as an agent, we may also attribute it mental attitudes. In [4] we observe three advantages of the attribution of mental attitudes to normative systems.

1. Obligations can be defined in the BDI framework. The desires or goals of the normative system are the obligations of the agent. This contributes to the open problem whether norms and obligations should be represented explicitly, for example in a deontic logic, or they can also be represented implicitly.
2. The interaction between an agent and the normative system can be modeled as a game between two agents. Consequently, methods and tools used in game theory such as equilibrium analysis can be applied to normative reasoning.
3. Behavior which counts as a violation is distinguished from behavior that is sanctioned. The normative system may autonomously decide which behavior counts as a violation, and whether violations are sanctioned.

#### 3.4 Comparison with artificial social systems

The characterization of social laws in classical game theory is more abstract than models of norm creation in normative multi-agent systems and deontic logic. However, this does not exclude the applicability of the ideas developed in normative multi-agent systems and deontic logic in artificial social systems. The assumption of this paper is that artificial social systems and normative multi-agent systems are studying a related problem, and that social laws and social norms are comparable mechanisms.

However, there is some terminology which differs in these two areas. For example, conventions are defined in artificial social systems as social laws in which there is only one strategy for each agent, whereas in normative multi-agent systems conventions are defined as norms which do not need a control system. These distinctions in terminology seem to refer to superficial distinctions only.

In the following section we show how, as suggested by the social delegation cycle, enforceable social laws can be defined in artificial social systems by introducing a control system. For example, the creation of enforceable social laws can be defined by first off-line designing a social law, then adding a control system such that the social law is stable, and finally checking that the stable social law with its associated control system is acceptable for the agents.

## 4. CONTROL SYSTEM FOR SOCIAL LAWS

In this section we bridge artificial social systems with normative multi-agent systems. In particular, we extend artificial social systems with a control system represented by a normative system. In this section, a control system is represented by a (control) strategy.

### 4.1 Normative game

Following Boella and Lesmo [3], we introduce a normative agent representing the normative system, which we call agent 0. We therefore have to extend the notion of a game from a two player game to a three player game. Strategies of agent 0 are called control strategies.

**DEFINITION 7.** A normative game (or a normative multi-agent encounter) is a three player game  $\langle N, R, S, T, U_1, U_2 \rangle$ , where  $N = \{0, 1, 2\}$  is a set of agents including agent  $0 \in N$  that represents the normative system,  $R, S$  and  $T$  are the sets of strategies available to agents 0, 1 and 2 respectively, and  $U_1 : R \times S \times T \rightarrow \mathbb{R}$  and  $U_2 : R \times S \times T \rightarrow \mathbb{R}$  are utility functions for agents 1 and 2, respectively.

### 4.2 Directly enforceable social laws

In this section we interpret the strategies of agent 0 as either introducing no control system, or selecting a control system from a set of alternatives. The state before and after the creation of a social law is modelled as a sub-game of the agents given by a strategy of agent 0. Given the strategies of agents 1 and 2, the outcome of the game is completely determined.

**DEFINITION 8.** Given a normative game represented by  $g = \langle \{0, 1, 2\}, R, S, T, U'_1, U'_2 \rangle$ , the sub-game of  $g$  defined by strategy  $r \in R$  is the game  $\langle \{1, 2\}, S, T, U_1, U_2 \rangle$  is the sub-game of the if and only if  $\forall s \in S, t \in T : U_1(s, t) = U'_1(r, s, t)$  and  $U_2(s, t) = U'_2(r, s, t)$ .

A social law is directly enforceable if there is a strategy for agent 0 such that  $\bar{S}, \bar{T}$  is a quasi-stable social law. We use quasi-stable social laws instead of stable social laws, because we do not want the social laws to restrict the freedom of the agents more than necessary. This allows the agents the freedom to act autonomously. We further address this issue when we discuss the choice of control systems in Section 6.

**DEFINITION 9.** Given a normative game represented by  $g = \langle N, R, S, T, U'_1, U'_2 \rangle$ , and an efficiency parameter  $e$ , A social law  $\bar{S}, \bar{T}$  (i.e., a restriction of  $S$  to  $\bar{S} \subseteq S$ , and of  $T$  to  $\bar{T} \subseteq T$ ) is directly enforceable if there is a strategy for agent 0 such that the social law is stable in the sub-game defined by this strategy.

An alternative and stronger definition of enforceable social laws asks not only for the existence of a quasi-stable social law, but it also asks that such a quasi-stable social law is unique. This represents that the control system identifies a unique solution to the agent's coordination problem. The following example illustrates the use of uniqueness in Lewis' classical coordination game, also discussed extensively by Tennenholtz.

**EXAMPLE 1.** Consider the game in Table 1. These tables should be read as follows. Strategies are represented by literals, i.e., atomic propositions or their negations. Each table

represents the sub-game given a strategy of agent 0, represented by  $\neg n$  and  $n$ , respectively. Agent 1 is playing columns and agent 2 is playing rows. The values in the tables represent the utilities of agent 1 and 2.

$\neg n$	$p$	$\neg p$	$n$	$p$	$\neg p$
$q$	1, 1	0, 0	$q$	1, 1	0, 0
$\neg q$	0, 0	1, 1	$\neg q$	0, 0	-1, -1

**Table 1: Control system for coordination game**

When the normative system plays  $\neg n$ , then agent 1 and 2 play a classical coordination game, which represents that the agents have to coordinate their actions to obtain the highest payoff. Agent 0 (the normative system) can play strategy  $\neg n$  or  $n$ , agent 1 can play strategy  $p$  or  $\neg p$ , agent 2 can play strategy  $q$  or  $\neg q$ . Intuitively, the strategy  $\neg n$  corresponds to the state before the social law is introduced, and  $n$  corresponds to the introduction of a control system that punishes an agent for deviating from  $p, q$ . When agent 0 plays  $\neg n$ , both  $p, q$  and  $\neg p, \neg q$  are stable social laws. However, when agent 0 plays  $n$ , then one of the stable social laws is eliminated, and only  $p, q$  is a stable social law.

### 4.3 Computational problem

The new computational problem is called the directly enforceable social law problem or DESLP.

**DEFINITION 10 (DESLP).** Given a normative game  $g$ , and an efficiency parameter  $e$ , find a Directly Enforceable Social Law which guarantees to the agents a payoff which is greater than or equal to  $e$  if such a law exists, and otherwise announce that no such law exists.

The condition that a social law is acceptable for the agents, is represented by setting the efficiency parameter  $e$  to a value higher than the expected pay-off before the social law is created. Directly enforceable social laws are illustrated by the following discussion of the prisoner's dilemma.

**EXAMPLE 2.** Consider the game in Table 2. When the normative system plays  $\neg n$ , the sub-game of agent 1 and 2 is a classical prisoner's dilemma. When the normative system plays  $n$ , the agents are never better off compared to the normative agent playing  $\neg n$ . Nevertheless, due to the dynamics of the game, the overall outcome is better for both agents. For example, in the sub-game defined by strategy  $\neg n$ , the only Nash equilibrium is 2, 2. Now suppose we set the efficiency parameter to 3, which means that all agents will be better off. If the normative system plays  $n$ , then the sub-game has a Nash equilibrium which is the (Pareto optimal) 3, 3. This explains why the agents accept the possibility to be sanctioned.

$\neg n$	$p$	$\neg p$	$n$	$p$	$\neg p$
$q$	3, 3	4, 1	$q$	3, 3	2, 1
$\neg q$	1, 4	2, 2	$\neg q$	1, 2	0, 0

**Table 2: Control system for prisoner's dilemma**

## 5. MONITORING VIOLATIONS

When not all violations are sanctioned, we may represent the behavior of the control system by a set of strategies (a kind of mixed strategy). In this section we characterize the monitoring of violations as a game involving the normative system.

### 5.1 Enforceable social laws

We first define maximin strategies in case agent 0 can choose from a set of strategies.

DEFINITION 11. Let  $R$ ,  $S$  and  $T$  be the sets of strategies available to agent 0, 1 and 2, respectively, and let  $U_i$  be the utility function of agent  $i$ .  $U_1(R, s, T) = \min_{r \in R, t \in T} U_1(r, s, t)$  for  $s \in S$ , and  $U_2(R, S, t) = \min_{r \in S, s \in S} U_2(r, s, t)$  for  $t \in T$ . The maximin value for agent 1 (respectively 2) is defined by  $\max_{s \in S} U_1(R, s, T)$  (respectively  $\max_{t \in T} U_2(R, S, t)$ ).

Again, a social law is useful with respect to an efficiency parameter  $e$  if each agent can choose a strategy that guarantees it a payoff of at least  $q$ .

DEFINITION 12. Given a normative game represented by  $g = \langle N, R, S, T, U_1, U_2 \rangle$  and an efficiency parameter  $e$ , we define a social law to be a restriction of  $S$  to  $\bar{S} \subseteq S$ , and of  $T$  to  $\bar{T} \subseteq T$ . The social law is useful if the following holds: there exists  $s \in \bar{S}$  such that  $U_1(R, s, \bar{T}) \geq e$ , and there exists  $t \in \bar{T}$  such that  $U_2(R, \bar{S}, t) \geq e$ .

We adapt the definition of quasi-stable social law to take sets of strategies into account. A social law is quasi-stable if an agent does not profit from violating the law, as long as the other agent conforms to the social law (i.e., selects strategies allowed by the law).

DEFINITION 13. Given a normative game represented by  $g = \langle N, R, S, T, U_1, U_2 \rangle$ , and an efficiency parameter  $e$ , a quasi-stable social law is a useful social law (with respect to  $q$ ) which restricts  $S$  to  $\bar{S}$  and  $T$  to  $\bar{T}$ , and satisfies the following: there is no  $s' \in S \setminus \bar{S}$  which satisfies condition  $U_1(R, s', \bar{T}) > \max_{s \in \bar{S}} U_1(R, s, \bar{T})$ , and there is no  $t' \in T \setminus \bar{T}$  which satisfies  $U_2(R, \bar{S}, t') > \max_{t \in \bar{T}} U_2(R, \bar{S}, t)$ .

We define enforceable social laws analogous to directly enforceable social laws.

DEFINITION 14. Given a normative game represented by  $g = \langle N, R, S, T, U_1, U_2 \rangle$ , and an efficiency parameter  $q$ , a social law (i.e., a restriction of  $S$  to  $\bar{S} \subseteq S$ , and of  $T$  to  $\bar{T} \subseteq T$ ) is enforceable if there is a restriction of  $R$  to  $\bar{R} \subseteq R$  such that  $\bar{S}, \bar{T}$  is quasi-stable in the normative game  $g = \langle N, \bar{R}, S, T, U_1, U_2 \rangle$ .

### 5.2 Computational problem

Replacing directly enforceable by enforceable social laws gives us the main problem.

$\neg n_1, \neg n_2$	$p$	$\neg p$	$n_1$	$p$	$\neg p$	$n_2$	$p$	$\neg p$
$q$	3,3	4,1	$q$	3,3	1,1	$q$	3,3	4,1
$\neg q$	1,4	2,2	$\neg q$	1,4	0,0	$\neg q$	1,1	0,0

Table 3: Mixed strategy of normative system

DEFINITION 15 (ESLP). Given a normative multi-agent encounter  $g$ , and an efficiency parameter  $e$ , find an Enforceable Social Law which guarantees to the agents a payoff which is greater than or equal to  $e$  if such a law exists, otherwise announce that no such law exists.

The monitoring of violations using enforceable social laws is illustrated by the following example.

EXAMPLE 3. Consider the game in Table 3. The first table represents that the normative system does not impose a sanctioning system, the second table represents that the first agent is being monitored and if he deviates from the law he is sanctioned, and the third table represents that the second agent is being monitored and thus in case of violation is sanctioned.

### 5.3 Sanction-based control systems

In this section we consider some additional assumptions on normative multi-agent encounters that reflect the concept of sanction-based obligations.

First, we assume that there is an ideal situation in which no norms are violated.

Secondly, we assume that the normative system only sanctions violations, while it does not reward good behavior. As a consequence of the first assumption, there is a strategy of the normative system in which the system does not sanction, and the payoffs in the ideal situation are higher than the payoffs in other situations. The game in Table 2 satisfies this assumption. The game in Table 4 does not, because the payoffs in the sub-game of strategy  $\neg n$  are higher for  $\neg p, q$ , and the payoffs in the sub-game of strategy  $n$  are higher for pay-off  $\neg p, \neg q$ .

$\neg n$	$p$	$\neg p$	$n$	$p$	$\neg p$
$q$	3,3	4,1	$q$	3,3	2,2
$\neg q$	1,4	2,2	$\neg q$	2,2	1,1

Table 4: Sanctions and rewards

Thirdly, we assume that the only agents sanctioned in a state are agents that have chosen a strategy which deviates from the ideal one. This represents the property that norms are individualized. The games in Table 2 and Table 4 satisfy this condition. The game in Table 5 does not, because the strategies  $\neg p, q$  lead to a decrease in pay-off not only for agent 1, but also for agent 2.

$\neg n$	$p$	$\neg p$	$n$	$p$	$\neg p$
$q$	3,3	4,1	$q$	3,3	0,0
$\neg q$	1,4	2,2	$\neg q$	0,0	0,0

Table 5: Sanctioning the innocent

Fourthly, the normative system only sanctions in case of violation, that is, it does not sanction good behavior. This implies that the control system does not come with a cost.

$\neg n$	$p$	$\neg p$	$n$	$p$	$\neg p$
$q$	3,3	4,1	$q$	2,2	1,1
$\neg q$	1,4	2,2	$\neg q$	1,1	0,0

Table 6: Sanctioning good behavior

## 6. CHOOSING A CONTROL SYSTEM

An important issue within artificial social systems is the choice of a particular social law, when there are several available. For example, Fitoussi and Tennenholtz [10] distinguish two criteria to choose social laws called *minimal* and *simple* social laws. In our case of enforceable social laws, there is an additional issue. Besides the choice for a social law, there is also for each social law a possible choice among control systems.

### 6.1 Minimal change

Consider the game in Table 7. When the normative system plays  $\neg n_1, \neg n_2, \neg n_3$ , then the agent 1 and 2 play again a classical coordination game. Now there are three control systems. The first control system  $n_1$  enforces  $p, q$  as the social law, and  $n_2$  enforces  $\neg p, \neg q$  as the social law. The choice between  $n_1$  and  $n_2$  chooses which social law is enforced (e.g., whether to drive on right or left side of the street). Strategy  $n_3$  is another control system which also enforces  $p, q$  as a social law, like  $n_1$ , but it does so using another control system, with other payoffs (e.g., whether people driving on the wrong side of street have to pay a penalty, or they are put in prison).

$\neg n_1, \neg n_2, \neg n_3$	$p$	$\neg p$	$n_1$	$p$	$\neg p$
$q$	1, 1	0, 0	$q$	1, 1	0, 0
$\neg q$	0, 0	1, 1	$\neg q$	0, 0	-1, -1
$n_2$	$p$	$\neg p$	$n_3$	$p$	$\neg p$
$q$	-1, -1	0, 0	$q$	1, 1	-1, -1
$\neg q$	0, 0	1, 1	$\neg q$	-1, -1	-2, -2

Table 7: Choice among three control systems

According to the minimal change criterium, control strategy  $n_1$  is preferred to control strategy  $n_3$ . A motivation for this criterium is that it may lead to simpler control systems.

### 6.2 Autonomy of agents

Freedom to act autonomously can be defined as set of strategies allowed by the social law. This criterium prefers control strategies which maximize the set of strategies of the enforced social law. Table 8 illustrates this criterium. Control strategy  $n_1$  does not restrict more than necessary, and leaves more freedom to act autonomously than control strategy  $n_2$ .

$\neg n_1, \neg n_2$	$p_1$	$p_2$	$\neg p_1, \neg p_2$
$q_1$	3, 3	3, 3	4, 1
$q_2$	3, 3	3, 3	4, 1
$\neg q_1, \neg q_2$	1, 4	1, 4	2, 2
$n_1$	$p_1$	$p_2$	$\neg p_1, \neg p_2$
$q_1$	3, 3	3, 3	1, 1
$q_2$	3, 3	3, 3	1, 1
$\neg q_1, \neg q_2$	1, 1	1, 1	0, 0
$n_2$	$p_1$	$p_2$	$\neg p_1, \neg p_2$
$q_1$	3, 3	0, 0	0, 0
$q_2$	0, 0	0, 0	0, 0
$\neg q_1, \neg q_2$	0, 0	0, 0	0, 0

Table 8:  $n_1$  leaves more freedom than  $n_2$

An additional advantage of leaving much freedom to the agents, is that in the future there is also more possibilities

to further restrict the agents' sets of strategies, that is, to introduce new social laws. For example, creation of the first social law is the identification of  $\bar{R} \subseteq R$ ,  $\bar{S} \subseteq S$  and  $\bar{T} \subseteq T$ , and the second social law creation is the identification of  $\underline{\bar{R}} \subseteq \bar{R} \subseteq R$ ,  $\underline{\bar{S}} \subseteq \bar{S} \subseteq S$  and  $\underline{\bar{T}} \subseteq \bar{T} \subseteq T$ . Example 4 illustrates this.

EXAMPLE 4. Consider the normative game in Table 9. The first table represents that the normative system does not impose a sanctioning system, the second table represents that there is a sanction for playing  $\neg p_1, \neg p_2$ , and the third table represents that there is an additional sanction for playing  $p_2$ . The first social law is  $\{p_1, p_2\}, \{q_1, q_2\}$  based on control system  $\{n_1, n_2\}$ , and the second social law is  $p_1, q_1$  based on control system  $n_1$ .

$\neg n_1, \neg n_2$	$p_1$	$p_2$	$\neg p_1, \neg p_2$
$q_1$	3, 3	4, 1	6, 0
$q_2$	1, 4	2, 2	0, 0
$\neg q_1, \neg q_2$	0, 6	0, 0	1, 1
$n_1$	$p_1$	$p_2$	$\neg p_1, \neg p_2$
$q_1$	3, 3	4, 1	0, 0
$q_2$	1, 4	2, 2	0, 0
$\neg q_1, \neg q_2$	0, 0	0, 0	0, 0
$n_2$	$p_1$	$p_2$	$\neg p_1, \neg p_2$
$q_1$	3, 3	1, 1	0, 0
$q_2$	1, 1	0, 0	0, 0
$\neg q_1, \neg q_2$	0, 0	0, 0	0, 0

Table 9: Iterated social law creations

### 6.3 Cost of control system

Another important criterium is to minimize the costs of the control system, as illustrated by the following table.

$\neg n_1, \neg n_2$	$p$	$\neg p$	$n_1$	$p$	$\neg p$	$n_2$	$p$	$\neg p$
$q$	3, 3	4, 1	$q$	3, 3	1, 1	$q$	1, 1	0, 0
$\neg q$	1, 4	2, 2	$\neg q$	1, 1	0, 0	$\neg q$	0, 0	0, 0

Table 10:  $n_2$  is more expensive than  $n_1$

### 6.4 Other criteria

Many other criteria can be defined. For example, we may prefer stable over quasi-stable social laws. Note that this criterium is not compatible with the autonomy criterium. Another criterium prefers unique social laws. A challenging example is discussed by Beccaria [2], who argues that it is not efficient to sanction minor offences with high penalties, because once a minor offence has been committed, there is no motivation anymore for an offender to obey the law. For example, if there is a death penalty for theft, then there is no motivation for thieves not to kill all witnesses. There are many aspects to Beccaria's example. First, there is the notion of monitoring: the thief has not been captured yet, and is playing a game such that if he is caught, he does not care whether he is caught for being a thief or for being a murderer. Secondly, another aspect of the example is that the thief is not making safety level decisions. He deliberately takes his chances of being caught as a thief and getting the death penalty for it. For this reason, we leave this example for further research.

## 7. FURTHER RESEARCH

In this section we consider various further extensions of artificial social systems, based on theories developed in normative multiagent systems and deontic logic. For example, we can model the normative system as a full agent with a utility function, and we can use role based organizations to explain how role playing agents determine the behavior of the normative system.

### 7.1 The utility function of agent 0

We consider in [6] the extension of normative games with a utility function of agent 0, to represent the norms which are enforced. Since agent 0 is a socially constructed agent, in the sense of Searle [13], its utility function can be updated. In particular, the enforcement of a social law by  $\bar{R} \subseteq R$  is represented by giving  $\bar{R}$  strategies a high utility, and  $R \setminus \bar{R}$  strategies a low utility. Moreover, we go beyond the framework of enforceable social laws by varying the utility of agent 0 depending on the strategies played by the other agents, and by considering incremental updates of the utility function to represent the evolution of artificial social systems. Formally, we extend a normative game with a utility function  $U_0 : R \times S \times T \Rightarrow \mathbb{R}$ , we define  $U_0(r, S, T) = \min_{s \in S, t \in T} U_0(r, s, t)$  for  $r \in R$ , and we define useful and quasi-stable social laws in the obvious way. Enforced social laws are defined as follows.

**DEFINITION 16.** *Given a normative game represented by  $g = \langle N, R, S, T, U_1, U_2 \rangle$ , and an efficiency parameter  $q$ , a social law (i.e., a restriction of  $S$  to  $\bar{S} \subseteq S$ , and of  $T$  to  $\bar{T} \subseteq T$ ) is enforced if there is a unique restriction of  $R$  to  $\bar{R} \subseteq R$  such that  $\bar{R}, \bar{S}, \bar{T}$  is quasi-stable.*

The game in Table 11 illustrates that the computational problem to find quasi-stable laws corresponds in extended normative games to the identification of enforced social laws. The values in the tables represent the utilities of agent 0 (in italics), 1 and 2.

$\neg n$	<i>p</i>	$\neg p$	<i>n</i>	<i>p</i>	$\neg p$
<i>q</i>	<i>3,3,3</i>	<i>0,4,1</i>	<i>q</i>	<i>3,3,3</i>	<i>1,2,1</i>
$\neg q$	<i>0,1,4</i>	<i>1,2,2</i>	$\neg q$	<i>1,1,2</i>	<i>0,2,2</i>

**Table 11: What is the enforced social law?**

We also consider the problem, given a normative game, to define a new utility function for the normative system. The principle that we like to maintain as much as possible from the existing social laws can be represented by the use of the principle of minimal change.

### 7.2 Roles, organizations, contracts, rights, . . .

Despite the fact that we model the normative system as an autonomous agent, the behavior of the normative system is determined by agents playing a role in it. We can model also this aspect of normative systems. For example, we can extend the model of artificial social systems with another class of socially constructed agents – called roles – determining the behavior of the normative system. In particular, we replace the set of strategies of the normative system  $R$  by a set of set of strategies of the roles in the normative system, written as  $R_1, \dots, R_n$ . A strategy of the normative system corresponds to a strategy for each of its roles, which is represented by  $R = R_1 \times \dots \times R_n$ . The strategies of the

role are its possible behaviors associated with exercising the rights of the role. We can represent the various roles in the Trias Politica, one for counting behavior as a violation, one for sanctioning in case the first one counts behavior as a violation, etc.

The use of organizations could be useful to explain the interaction of social laws and other social concepts such as roles. Moreover new computational problems can be defined, such as an organizational design problem (decompose the organization into a set of roles such that the organizational goals are achieved if the roles' goals are achieved), a role assignment problem (assign real agents to roles such that goals of roles and thus goals of organization are achieved), etc. Organizational structures can also be used to introduce a multi-agent structure of the normative system, defining multiple normative systems motivating each other.

The interaction structure in many multiagent systems is not completely fixed in advance to preserve the autonomy of agents. For example, in (virtual) organizations the interaction possibilities can be changed and negotiated. For this reason, several approaches introduce the possibility for agents to stipulate contracts. A contract can be defined as a statement of intent that regulates behavior among organizations and individuals. Contracts have been proposed as means to make explicit how agents can change the interaction with and within the organization: they create obligations, permissions and new possibilities of interactions. From a contractual perspective, organizations can be seen as the possible sets of agreements for satisfying the diverse interests of self interested individuals.

Rights have been addressed by Alonso [1], who says that “a right is considered as a set of restrictions on the agent’s activities which allow them enough freedom, but at the same time constraint them.” He then continues to distinguish rights from social laws, and illustrates his notion by an example from traffic law, where two cars would drive into each other on a crossroads. However, we believe that this example should be modeled with obligations or prohibitions instead of rights, and in general that the characteristic property of a right is that it *increases* the set of possible agent strategies. Moreover, role assignment means that an agent can decide the strategy the role is playing (i.e., decide which rights are exercised). Of course, roles come not only with rights but only with responsibilities and obligations that may decrease the agent’s freedom.

### 7.3 Game theory

Of course, not only insights of normative multi-agent systems and deontic logic can be used for further developing artificial social systems. There are many issues discussed in game theory which are relevant for the further development of artificial social systems. For example, we can replace the present classical game-theoretic setting of artificial social systems, where everything is known to every agent, by variants of game theory that include uncertainty.

Stability of artificial social systems is defined as following a norm when other agents do so. However, in game theory the following situation has been discussed. Two agents do not want to meet. Agent 1 can go to room a or room b, agent 2 can go to room b or room c. The equilibrium is that agent 1 goes to room a and agent 2 goes to room c. However, this is not stable in the sense that in this state, agent 1 can as well go to room b.

## 8. SUMMARY

In this paper we extend the artificial social systems approach with the notion of enforceable social laws. We introduce the enforceable social law problem for both directly enforceable social laws, and social laws with a monitoring system. This new computational problem is based on a bridge between two important theories of social systems. On the one hand artificial social systems based on social laws, and on the other hand normative multi-agent systems based on norms and deontic logic.

In this new setting of artificial social systems, we can distinguish between the choice of social law (that is, the sets of strategies the agents can play according to the social law), and the choice among control systems. For example, we can first design a social law and therefore design a control system, but we may also do so in parallel.

Having established the bridge between artificial social systems and normative multi-agent systems, we can start to use results and techniques from one area into the other one. In particular:

**Artificial social systems** can be further extended using the concepts developed in normative multi-agent systems, as discussed in Section 7.

**Normative multi-agent systems** can be studied from a more computational viewpoint. Moreover, the emergence of social norms can be studied following the emergence of social laws. In this setting, also the sanctions themselves can emerge.

Other issues for further research are determining the complexity of the new computational problems, and defining new computational problems for the choice of control system.

## 9. REFERENCES

- [1] E. Alonso. Rights and argumentation in open multi-agent systems. *Artificial Intelligence Review*, 21:3–24, 2004.
- [2] C. Beccaria. *Dei delitti e delle pene*. Livorno, 1764.
- [3] G. Boella and L. Lesmo. A game theoretic approach to norms. *Cognitive Science Quarterly*, pages 492–512, 2002.
- [4] G. Boella and L. van der Torre. Attributing mental attitudes to normative systems. In *Procs. of AAMAS'03*, pages 942–943. ACM Press, 2003.
- [5] G. Boella and L. van der Torre.  $\Delta$ : The social delegation cycle. In *LNAI n.3065: Procs. of  $\Delta EON'04$* , pages 29–42, Berlin, 2004.
- [6] G. Boella and L. van der Torre. The evolution of artificial social systems. In *Procs. of IJCAI'05*, 2005.
- [7] R. Brafman and M. Tennenholtz. Efficient learning equilibrium. *Artificial Intelligence*, 159(1-2):27–47, 2004.
- [8] W. Briggs and D. Cook. Flexible social laws. In *Procs of IJCAI'95*, pages 688–693. Morgan Kaufmann, 1995.
- [9] C. Castelfranchi, F. Dignum, C. Jonker, and J. Treur. Deliberate normative agents: Principles and architecture. In *Intelligent Agents VI (ATAL'99)*, Springer Verlag, 2000.
- [10] D. Fitoussi and M. Tennenholtz. Choosing social laws for multi-agent systems: Minimality and simplicity. *Artificial Intelligence*, 119(1-2):61–101, 2000.
- [11] E. Goffman. *Strategic interaction*. Basil Blackwell, Oxford, 1970.
- [12] A. Jones and J. Carmo. Deontic logic and contrary-to-duties. In D. Gabbay, editor, *Handbook of Philosophical Logic*, pages 203–279. Kluwer, 2002.
- [13] J. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.
- [14] Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73 (1-2):231 – 252, 1995.
- [15] Y. Shoham and M. Tennenholtz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94 (1-2):139 – 166, 1997.
- [16] M. Tennenholtz. On stable social laws and qualitative equilibria. *Artificial Intelligence*, 102 (1):1–20, 1998.
- [17] R. Wieringa and J.-J. Meyer. Applications of deontic logic in computer science: A concise overview. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 17–40. John Wiley & Sons, Chichester, England, 1993.