On Stable Social Laws and Qualitative Equilibrium for Risk-Averse Agents

Moshe Tennenholtz
Faculty of Industrial Engineering and Management
Technion–Israel Institute of Technology
Haifa 32000
Israel
e-mail: moshet@ie.technion.ac.il

Abstract

This paper introduces and investigates the notion of qualitative equilibria, or *stable social laws*, in the context of qualitative decision making. Previous work in qualitative decision theory has used the *maximin* decision criterion for modelling qualitative decision making. When several decision-makers share a common environment, a corresponding notion of equilibrium can be defined. This notion can be associated with the concept of a *stable social law.* This paper initiates a basic study of stable social laws; in particular, it discusses the stability benefits one obtains from using social laws rather than simple conventions, the existence of stable social laws under various assumptions, the computation of stable social laws, and the representation of stable social laws in a graph-theoretic framework.

1 Introduction

General coordination mechanisms are essential tools for efficient reasoning in multi-agent AI systems. Coordination mechanisms are a major issue of study in the fields of mathematical economics and game theory as well. Much work in these fields concentrates on the notion of an *equilibrium*. An equilibrium is a joint behavior of agents, where it is irrational for each agent to deviate from that behavior. The notion of an equilibrium discussed in the game theory and mathematical economics literature refers to agents which are expected utility maximizers. However, much work in AI has been concerned with more qualitative forms of rational decision making. In particular, work in AI has been concerned with agents which attempt to maximize their worst case payoff. Although, at first, this behavior may look questionable from a decision-theoretic perspective, it is known to capture the behavior of risk-averse agents (Luce & Raiffa 1957; Dubois & Prade 1995; Brafman & Tennenholtz 1996), and it is appropriate in the context of qualitative decision theory (Boutilier 1994; Tan & Pearl 1994; Dubois & Prade 1995; Darwiche & Goldszmidt 1994). Moreover, in (Brafman & Tennenholtz 1996) Brafman and Tennenholtz have shown general conditions under which an agent can be viewed *as if* it were a *maximin* agent (i.e., an agent which maximizes its worst case payoff). However, the corresponding notion of equilibrium has not yet been investigated. In this paper we introduce this notion and investigate its properties. The concept of qualitative equilibrium for risk-averse agents turns out to coincide with the notion of a *stable social law*, to be introduced later in this paper. For ease of exposition we introduce the notion of a stable social law in a self-contained fashion, as an extension to previous work on *artificial social systems.*

Some work on multi-agent systems assumes that agents are controlled by a single entity which dictates their behavior at each point in time, while some other work is concerned with decentralized systems where no global controller exists. A significant part of the theory developed for decentralized multi-agent systems (Bond & Gasser 1988; Demazeau & Muller 1990) deals with conflict resolution in multi-agent encounters. The basic theme of work on this subject is that in decentralized systems agents will reach states of conflict and appropriate negotiation mechanisms would be needed in order to resolve these conflicts. The result of the negotiation process is a deal that the agents will follow. Work in AI has been mostly concerned with agents that conform to agreed-upon deals. Agents may not follow irrational negotiation protocols, but will conform to deals obtained by following rational negotiation protocols (Kraus & Wilkenfeld 1991; Zlotkin & Rosenschein 1993; Durfee, Lee, & Gmytrasiewicz

1993).[1] This differs from work in game-theory (Owen 1982; Fudenberg & Tirole 1991) where a joint strategy is considered unstable (and therefore unsatisfactory from a design perspective) if an agent has a rational incentive to deviate from it. The Artificial Social Systems approach (e.g., (Moses & Tennenholtz 1990; Shoham & Tennenholtz 1995)) exposes a spectrum between a totally centralized approach and a totally decentralized approach to coordination. 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. In a mobile robots setting, for example, such a social law may consist of various traffic constraints (Shoham & Tennenholtz 1992). More generally, a social law is a set of *restrictions* on the agents' activities which allow them enough freedom on one hand, but at the same time constrain them so that they will not interfere with each other. In particular, a social law makes certain conflicts unreachable, and as a result improves the system efficiency. Notice that mechanisms for conflict resolution can serve as part of the social law; they will be used in situations where conflicts can't be prevented in advance.

The motivation for the theory of artificial social systems has been the design of artificial multi-agent systems, and as such it assumes that the agents will obey the law supplied by the designer. However, if each agent is designed by a different designer then some laws might be considered irrational. Therefore, at the current stage, the artificial social systems approach and approaches to conflict resolution are somewhat complementary; the resolution of conflicts in multi-agent encounters is part of a more general theory of social laws, but the theory of artificial social systems has neglected the stability of social laws in multi-agent encounters. In this paper we wish to bridge part of the gap between the theory of artificial social systems and the theory of conflict resolution in multi-agent encounters, by considering *stable social laws* for multi-agent encounters. A social law for a multi-agent encounter is a restriction of the set of available actions (in the encounter) to a set of socially allowed actions. Stable social laws make deviation from them irrational. Notice that a convention is a particular type of a social law; a convention determines a particular joint action for the agents to follow (e.g., keep the right of the road), while a social law allows several such actions and prohibits others. As it turns out, this distinction is quite important and useful in the non-bayesian context frequently adopted in AI.

[1] See (Sandholm & Lesser 1995) for a detailed discussion of this point.

We will discuss social laws for multi-agent encounters using a game-theoretic framework which is tailored for assumptions made in the AI literature, and especially in recent work on qualitative decision making. In particular, we will assume that the agents are risk-averse agents, which use the *maximin decision criterion*. More specifically, given a set of possible behaviors of the other agents, the aim of an agent is to optimize its worst case outcome assuming the other agents may follow any of these behaviors. This kind of behavior is appropriate where there is some ordinal relation on possible outcomes. In such situations all that matters to agents is the order of payoffs and not their exact value. The precise conditions under which such modelling is an appropriate one are discussed in (Brafman & Tennenholtz 1996). We will require that a social law suggested for a particular encounter will guarantee to the agents a certain payoff, and that it will be stable; there should be no incentive to deviate from it assuming the agents are risk-averse agents. Hence, a stable social law corresponds to a notion of qualitative equilibrium for risk-averse agents.

We start by introducing our framework. In particular, we define the notion of stable social laws. Having the basic framework, we show that the set of multi-agent encounters for which a stable convention exists is a strict subset of the set of multi-agent encounters for which there is an appropriate stable social law; however, we show that there exists situations where no stable social law exists. Then, we initiate a computational study of stable social laws; we formulate the corresponding computational problem and show that the general problem of coming up with a stable social law is intractable; in addition, we point to an interesting restriction on our framework under which the synthesis of stable social laws is polynomial. We then return back to the question of the existence of stable social laws; we first show how this question can be formulated in standard graph-theoretic terms, and then expose a class of encounters where simple graph-theoretic conditions imply the existence of stable social laws. Sketch of proofs can be found in the appendix.

2 The Basic Framework

In this section we introduce our basic framework, which is built upon a basic game-theoretic model.

2.1 The Basic Model

In general AI planning systems, agents are assumed to perform *conditional plans.*[2] A conditional plan is a (perhaps partial) function from the local state of an agent to action. Conditional plans can be treated as *protocols* in distributed systems terms, or as *strategies* in game-theoretic terms. In the sequel we will make use of a game-theoretic model; therefore, we will adopt the term strategy.

Multi-agent encounters can be represented as a game. In this paper we will consider two-person games, where two agents participate in an encounter. We will be concerned with finite games where each agent has a finite number of strategies. A *joint strategy* for the agents consists of a pair of strategies, one for each agent. Each joint strategy is associated with a certain payoff for each of the agents, as determined by their *utility functions.* The above-mentioned terms are classical game-theoretic terms which capture general multi-agent encounters.

Formally, we have:

Definition 2.1:

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 $U_1 : S \times T \rightarrow \Re$ and $U_2 : S \times T \rightarrow \Re$ are utility functions for agents 1 and 2 respectively.

One interesting point refers to the knowledge of the agents about the structure of the game. In this work, we assume that agents are familiar with the sets of actions available to the different agents, but an agent might be aware only of its own payoff function. Our results are appropriate both for the case where the payoff functions are common-knowledge among the agents and for the case an agent knows only its individual payoff function.

How should agents behave in a multi-agent encounter prescribed by a given game? The system's designer may wish to guarantee to the agents a particular payoff. Methods of negotiation for guaranteeing particular types of efficient behavior are discussed in the Distributed AI literature. Most of this literature assumes that agents do not deviate from agreed-upon joint strategies, although agents may adopt only rational negotiation protocols (Zlotkin & Rosenschein 1993; Kraus & Wilkenfeld 1991). On the other hand, work in Game-Theory has been concerned with finding joint strategies which will be stable against rational deviations, where rationality is associated with expected utility maximization. We are interested in guaranteeing efficient behavior for the agents; this behavior should be stable against rational deviations; however, the notion of rationality we adopt would be different from expected utility maximization and would be in the spirit of qualitative decision theory (Luce & Raiffa 1957; Dubois & Prade 1995; Brafman & Tennenholtz 1996). In this paper we borrow a most classical decision criterion in order to model a rational decision by the agent. Namely, we use the *maximin* decision criterion to be discussed below. The precise conditions under which an agent can be viewed as if it uses this decision criterion are discussed in (Brafman & Tennenholtz 1996). This modeling perspective becomes especially appealing if an agent does not know the payoff function of the other agent.

Definition 2.2: Let S_i be the set of strategies available to agent i, and let u_i be the utility function of agent i. Define $u_1(s, S_2) = min_{t \in S_2} u_1(s, t)$ for $s \in S_1$, and $u_2(S_1, s) = min_{t \in S_1} u_2(t, s)$ for $s \in S_2$. The *maximin value for agent 1 (resp. 2)* is defined by $max_{s \in S_1} u_1(s, S_2)$ (resp. $max_{t \in S_2} u_2(S_1, t)$). A strategy of agent i leading to the corresponding maximin value is called a *maximin strategy for agent i.*

In the sequel, we will assume the agents adopt the maximin decision criterion (i.e., choose a maximin strategy), although many of our results and observations do hold for other qualitative decision criteria as well.

2.2 Conventions and Social Laws

Given a game and a requirement that the agents will be able to obtain a payoff of at least t, the designer may supply the agents with an appropriate convention: a joint strategy for which the utility for both agents is greater than or equals to t. A convention is a special case of a social law. A social law in a multi-agent encounter is a restriction on the set of strategies available to the agents; a convention will simply restrict the behavior to a one particular joint strategy. When selecting a social law the designer may wish to select it in a way which allows each agent at least one strategy which guarantees a payoff of at least t.

Definition 2.3: Given a game $g = \langle N, S, T, U_1, U_2 \rangle$ and an efficiency parameter t, we define a *useful social law* to be a restriction of S to $\bar{S} \subseteq S$, and of T to $\bar{T} \subseteq T$, which satisfies that there exists $s \in \bar{S}$ such

[2] Plans with complete information and other forms of plans will be taken as restrictions on the general form of plans considered in this paper; this point will not effect the discussion or results presented in this paper.

that $U_1(s, \bar{T}) \geq t$, and that there exists $k \in \bar{T}$ such that $U_2(\bar{S}, k) \geq t$. A (useful) convention is a (useful) social law where $|\bar{S}| = |\bar{T}| = 1$.

In general, a useful social law is a restriction on each agent's activities which enable each agent to act individually and succeed reasonably well, as long as all the agents conform to the law (see the discussion and the general semantics in (Moses & Tennenholtz 1995; Shoham & Tennenholtz 1995)). At this point the idea of using social laws for coordinating agents' activities in a multi-agent encounter may seem a bit strange; why should we care about social laws if every efficiency degree which can be obtained by a social law can already be obtained by an appropriate simple convention? However, as we will see later, social laws can serve as much more useful entities than simple conventions for agents participating in a multi-agent encounter. We will elaborate on this point later.

3 Stable Social Laws

The concept of a social law which has been discussed in previous work is a general and most powerful tool. In this work we are concerned with social laws for multi-agent encounters. Although that's a most popular setting for the study of the resolution of conflicts, up to date the power of social laws has been illustrated in more complex settings (Shoham & Tennenholtz 1995; 1992). As we shall see, social laws may serve as useful tools for multi-agent encounters as well.

Definition 3.1: Given a game $g = \langle N, S, T, U_1, U_2 \rangle$ and an efficiency parameter q, 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:[3] 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 $U_2(\bar{S}, t') > max_{t \in \bar{T}}\{U_2(\bar{S}, t)\}$.

Hence, 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 a specific goal to obtain, and there is no reason to assume an agent will execute a strategy which yields to it a payoff which is lower than the payoff guaranteed to it by another strategy, assuming the other agent obeys the law. Putting it in other terms, given that we talk about a specific encounter with specific goals, there is no reason to include in the set of allowed strategies a strategy which is dominated by another strategy in that set. This requirement is consistent with models of stable social situations discussed in the game theory literature (Greenberg 1990).

Formally, we have:

Definition 3.2: 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 (i.e., selects strategies allowed by the law).

Notice that a stable social law is the equilibrium concept which one may wish to associate with the *maximin* decision criterion. Hence, our study of stable social laws can be interpreted as a basic study of equilibria in the context of qualitative decision making.

In the rest of this paper we discuss stable social laws. Given a multi-agent encounter, a stable social law will guarantee to the agents a particular payoff, similarly to the way a particular payoff can be guaranteed by a simple convention. As it turns out however, the difference between social laws and conventions stems from the fact that social laws may be more stable than conventions in multi-agent encounters. This will be the topic of the following section.

4 Social Laws vs. Conventions

Having a definition of stable social laws, one may ask: what are these laws good for? If we wish to guarantee a certain payoff for the agents, why can't we look for stable conventions, i.e., select a joint strategy from which a deviation would be irrational, assuming such a strategy exists?

The answer is supplied by the following result:

Theorem 4.1: *There exists games for which there are no stable conventions, but where appropriate stable social laws do exist.*

The above theorem reveals a new contribution of the theory of social laws: restricting the activities of the agents to a set of allowed actions rather than to a particular action is useful even in simple multi-agent encounters. This is due to the fact that social laws may be more stable than simple conventions. The intuition behind the above result is as follows. Although different strategies may lead to similar payoffs, different strategies may block different deviations by the agents. Therefore, the fact that the agent's behavior

[3] Our definition is in the spirit of classical game-theory; we require that deviation by one agent will be irrational given that the other agent sticks to the suggested behavior.

is only partially defined may improve the system efficiency. Assume for example that there are two agents, each of which can invest its money using four options, A, B, C,or D. If they will invest only in options A and B then they will get reasonable payoffs. However, if they are told to invest in particular options, e.g., one is told to invest in A and the other is told to invest in B, then one of them may take this opportunity in order to gain more on behalf of the other using option C or D. But, if both C and D yield low payoffs when they are applied against A or B (although not against both of them), such deviation can be prevented by telling each agent to choose (non-deterministically) from among options A and B (i.e., by supplying the social law: "don't use C and D", rather than pointing to particular investments). The reader may get additional understanding of this situation by considering the proof of Theorem 4.1.

We have shown that social laws are more stable than conventions. We can also show:

Theorem 4.2: *There exist games for which no stable social laws exist (for any efficiency parameter).*

Hence, stable social laws are powerful but do not always exist. Given this observation, it may of interest to supply a procedure for computing when a social law exists. Naturally, in cases where a stable social law exists it may be of interest to compute such a law. In addition, it may be of interest to characterize conditions for the existence of stable social laws. These are the topics of the following sections.

5 Computing Stable Social Laws

In this section we take a look at the computation of stable social laws. In order to do so, we first need to decide on the representation of our input. We will use the standard game-theoretic representation in which a multi-agent encounter is represented by a game matrix.

The problem of computing a Stable Social Law (SSLP) is defined as follows:

Definition 5.1: The Stable Social Law Problem [SSLP]:

Given a multi-agent encounter g, and an efficiency parameter t, find a Stable Social Law which guarantees to the agents a payoff which is greater than or equals to t if such a law exists, and otherwise announce that no such law exists.

Notice that if we restrict ourselves to simple conventions, the computational problem is easy; however, as we have observed, conventions are not as useful as social laws. As the following Theorem shows this does not come without a cost. We are able to show:

Theorem 5.2: *The SSLP is NP-complete.*

The importance of the above theorem is twofold: first, it supplies an initial result in the computational study of stable social laws. Second, the proof of this result teaches us about the structure of stable social laws. Further understanding of this structure is obtained in the following section, where we supply a graph-theoretic representation of stable social laws.

Given the previous Theorem, it would be of interest to identify general cases where the problem of coming up with a stable social law (if such a law exists) is tractable. One case which is of interest is when the parties involved are of unequal power. One way of capturing this fact is by assuming that one party has much more strategies available to it than the other party does. Formally, we say that an agent is *logarithmically bounded* if the number of strategies available to it is $O(log(n))$ where n is the number of strategies available to the other agent. In this case we can show:

Theorem 5.3: *The SSLP when one of the agents is logarithmically bounded is polynomial.*

6 Graph-Theoretic Representations of Stable Social Laws

We can learn about the structure of stable social laws by studying the reduction in Theorem 5.2. More generally, the study of a new equilibrium concept and of its use can greatly benefit from representation theorems which show what does this concept mean in terms of known concepts. In addition, in the context of this particular work, such representation theorems can supply conditions for the existence of stable social laws. The reduction used in the proof of Theorem 5.2 shows that a *special case* of the problem of finding a stable social law is isomorphic to a well-known problem. This has been useful for proving the above-mentioned result. However, it would be of interest to characterize the general Stable Social Laws concept by means of well-known terminology. In particular, in this section we make use of graph-theoretic terms in order to characterize the stable social law concept.

We will make use of the following standard terms:

Definition 6.1: Let $G = (V, E)$ be a *graph*, where V is a set of *nodes*, and $E \subseteq V^2$ is a set of *edges*.

G is *undirected* if, for all $v_1, v_2 \in V$, $(v_1, v_2) \in E$ iff $(v_2, v_1) \in E$, and is *directed* otherwise. A set $V' \subseteq V$ is an *independent set* if there are no $v', v'' \in V'$ which satisfy $(v', v'') \in E$. A set $V' \subseteq V$ is a *clique* if $(v', v'') \in E$ for all $v', v'' \in V'$. A node $v \in V$ is *non-isolated* relative to $V' \subseteq V$ if there is a vertex $v' \in V'$ such that $(v, v') \in E$. A set $V' \subseteq V$ is a *dominating set* if for each node $v' \in V - V'$ there is a node $v'' \in V'$ such that $(v', v'') \in E$. A node $v \in V$ is a *sink* if there is no v' such that $(v, v') \in E$. The graph G is *k-colorable* if we can color the nodes of the graph with k colors in a way that $(v, v') \in E$ implies that v and v' have different colors.

We would now like to make a connection between the above-mentioned graph-theoretic terms and our notion of a stable social law. In the sequel we will be concerned with games where the sets of strategies, S, available to the agents are identical. We will also assume the game is symmetric in the sense that $U_1(s,t) = U_2(t,s)$ (i.e., the outcome of the agents is independent of their names). We will be interested in social laws that are fair, in the sense that if a strategy is prohibited for one agent then it is prohibited for all agents. For ease of exposition we will be concerned with social laws guaranteeing the value t and no more than t.

Definition 6.2: Given a game g and an efficiency parameter t, let $G_1 = (V, E_1), G_2 = (V, E_2), G_3 = (V, E_3)$ be directed graphs where V is associated with the set of strategies S, and E_i is defined as follows: $(s,q) \in E_1$ iff $U_1(s,q) \geq t$; $(s,q) \in E_2$ iff $U_1(s,q) = t$; $(s,q) \in E_3$ iff $U_1(s,q) \leq t$

Given the above-mentioned graphs which are built based on the game g and the efficiency parameter t, we can show the connection between stable social laws and standard graph-theoretic concepts:

Theorem 6.3: *Given a game g and an efficiency parameter t, the corresponding graphs G_1, G_2, G_3 satisfy the following: a stable social law for g exists iff there is a subset V' of the nodes of the related graphs, such that V' is a clique in G_1, a dominating set in G_3, and all nodes in V' are non-isolated, relative to V', in G_2.*

The above theorem supplies an additional graph-theoretic understanding of the notion of stable social laws. A further look at such representations enables us to prove additional general existence theorems for stable social laws. One interesting general type of multi-agent encounters refers to games which are a combination of pure coordination and zero-sum games. The importance of such type of games is obvious; they allow agents either to agree and obtain "reasonable payoff" or to "fight" for "high payoff" taking the risk of obtaining "low payoff". These basic games are formally defined as follows:

Definition 6.4: Assuming w.l.o.g that the efficiency parameter t equals 0, a symmetric game g is a *mixed coordination-competition game*, if the utility functions satisfy:

1. $U_1(s,s) = 0$ for every $s \in S$.
2. $U_1(s,q) > 0$ iff $U_1(q,s) < 0$ for every $s, q \in S$.

An interesting point about mixed coordination-competition games is that they can be represented by a single graph, $\bar{G}$, which is defined as follows: the nodes of $\bar{G} = (\bar{V}, \bar{E})$ corresponds to the different strategies, and the set of edges $\bar{E}$ is defined as follows: $(s,t) \in \bar{E}$ iff $U_1(s,t) < 0$. Given this graph structure, we can prove the existence of stable social laws for an interesting class of encounters:

Theorem 6.5 : *Given a mixed coordination-competition game g, if the corresponding graph $\bar{G}$ has a sink or is 2-colorable then an appropriate stable social law exists.*

7 Other Qualitative Equilibiria

The previous sections have been concerned with qualitative equilibrium, where the decision criterion is the *maximin* decision criterion. As we have mentioned before, our discussion and results do hold for other decision criterion as well. In this section we take a look at two other basic decision criretia, the *minimax regret* decision criretion, and the *competative ratio* decision criterion, and show how the previous study can be applied to the context of these decision criteria as well.

Definition 7.1 : Let S_i be the set of strategies available to agent i, and let u_i be the utility function of agent i. Given $s \in S_1$, and $q \in S_2$, define $u_1(s, q, S_2) = max_{t \in S_2} u_1(s,t) - u_1(s,q)$. Given $q \in S_1$ and $s \in S_2$ define $u_2(S_1, q, s) = max_{t \in S_1} u_2(t,s) - u_2(q,s)$. The *minimax regret value for agent 1 (resp. 2)* is defined by $min_{s \in S_1} max_{q \in S_2} u_1(s, q, S_2)$ (resp. $min_{t \in S_2} max_{q \in S_1} u_2(S_1, q, t)$). A strategy of agent i leading to the corresponding minimax value is called a *minimax strategy for agent i.*

8 Discussion

In this work we have introduced a theory of stable social laws, or qualitative equilibria, for risk-

averse agents. Our work bridges some of the gap between work on Artificial Social Systems and work on conflict resolution in Game Theory and AI. Social laws have been shown to be a basic and useful tool for the coordination of multi-agent systems (Moses & Tennenholtz 1990; Shoham & Tennenholtz 1995; Briggs & Cook 1995; Minsky 1991). However, the stability of social laws in a system of rational agents has been neglected so far. This work extends previous work on social laws for artificial agent societies by considering stable social laws for multi-agent encounters.

Two major lines of research related to our work are work in the field of Game Theory and work in the field of Distributed Artificial Intelligence [DAI]. Related work on rational deals and negotiations in DAI (e.g., (Rosenschein & Genesereth 1985; Zlotkin & Rosenschein 1993; Kraus & Wilkenfeld 1991)) have adopted a game-theoretic perspective. A very interesting property of this work is that it considers deals among rational agents who will not deviate from agreed-upon deals.[4] In difference to this assumption, our work is concerned with agents who will deviate from agreed-upon deals if they have a rational incentive to do so.

Much work in Game Theory has been concerned with devising rational conventions for a group of rational agents; a rational agent may deviate from a prescribed joint strategy if this deviation will improve its own situation. More specifically, much work in Game Theory (Luce & Raiffa 1957; Owen 1982; Fudenberg & Tirole 1991) has been devoted to the study of equilibrium in games; an equilibrium will have the property that there is no rational incentive for an agent to deviate from the equilibrium as long as other agents stick to it. The notion of an equilibrium has been adopted to the AI and DAI literature in various settings (e.g., (Wellman 1993)), as part of a general and important attempt to introduce social and organizational metaphors into the AI context (Simon 1981; Fox 1981; Malone 1987; Durfee, Lesser, & Corkill 1987; Doyle 1983; Davis & Smith 1983; Jennings 1995; Ishida, m. Yokoo, & Gasser 1990; Cohen & Levesque 1991; Gasser 1993). A central notion in this regard is the notion of a rational agent adopted from the decision/game theory literature. Most work in Game-Theory has associated the notion of a rational agent with the notion of expected utility maximization. This is not however the usual way a rational agent is viewed in the AI literature, such as in work on conditional planning (Warren 1976; Peot & Smith 1992; Etzioni *et al.* 1992; Genesereth & Nourbakhsh 1993; Safra & Tennenholtz 1994).

[4] This is not to say that other assumptions are not treated by the DAI literature; see for example (Sandholm & Lesser 1995).

Using a Game-Theoretic terminology, in this work we developed an equilibrium theory for risk-averse agents (Luce & Raiffa 1957; Dubois & Prade 1995; Brafman & Tennenholtz 1996). It turns out that the notion of stable social laws is a powerful tool in this regard. Consider general multi-agent encounters, the notion of social law seems to serve as a useful mechanism for obtaining stable situations for risk-averse agents, similarly to the way mixed strategies serve as useful tools for expected utility maximizers. Our theory and results can therefore be interpreted both as an extension to the theory of Social Laws presented in the AI literature, as well as a contribution to the foundations of discrete/qualitative Decision/Game Theory. We hope it can lead to further cross-fertilization between these fields.

Appendix: Sketch of Proofs

Proof of Theorem 4.1 (sketch):

The proof follows by considering the following game:

	agent 2			
agent 1	A	B	C	D
A	(1,1)	(1,1)	(2,0)	(0,2)
B	(1,1)	(1,1)	(0,2)	(2,0)
C	(2,0)	(0,2)	(0.5,0.5)	(0.75,0.25)
D	(0,2)	(2,0)	(0.25,0.75)	(0.5,0.5)

Assume the designer wishes to guarantee the payoff 1. The fact that no stable convention exists follows by case analysis. On the other hand, if we restrict both agents to perform actions taken from $\{A, B\}$ then a payoff of 1 is guaranteed for both of the agents and no deviation is rational. ∎

Proof of Theorem 4.2 (sketch):

The proof follows by considering the following game:

	agent 2	
agent 1	A	B
A	(2.5,1)	(1.5,3)
B	(2,2)	(4,0.5)

A case analysis shows that no stable social law exists in the above-mentioned game.

∎

Proof of Theorem 5.2 (sketch):

The proof that the problem is in NP is straightforward. The proof the problem is NP-hard is by resuction from 3-SAT (Garey & Johnson 1979). Given a 3-CNF formula φ we generate a game g, for which a stable social law exists if and only if φ is satisfiable. We take the efficiency parameter t to be equal to 0, and let t' be a positive real number.

With clause number i in φ we associate the strategies $c_i^1, c_i^2, \ldots, c_i^7$ and d_i; each $c_i{}^k$ is associated with a different truth assignment to clause i (there are seven such assignments), and d_i is an additional distinguished strategy which is associated with that clause. The set of strategies for each player in the game g is the union of all strategies which are associated with the different clauses in φ.

We take g to be a symmetric game, and specify the utility function of agent 1:

1. $U_1(d_i, d_j) = -t'$ for all i, j.
2. $U_1(c_i^k, c_j^l) = 0$ iff c_i^k and c_j^l correspond to consistent assignments.
3. $U_1(c_i^k, c_j^l) = -t'$ iff c_i^k and c_j^l correspond to inconsistent assignments.
4. $U_1(d_i, c_j^k) = t'$ for all $i \neq j$ and every k.
5. $U_1(c_j^k, d_i) = -t'$ for all $i \neq j$ and every k.
6. $U_1(d_i, c_i^k) = -t'$ for every i and every k.
7. $U_1(c_i^k, d_i) = t'$ for every i and every k.

Now, consider a truth assignment T which satisfies φ. We can define a social law which leaves each agent only with the strategies which their corresponding assignments are as determined by T (and with no strategy of the form d_i). It is easy to see that we get a stable social law; the social law guarantees a payoff of 0 since the agents are left only with "consistent strategies", and deviations are irrational since there is a representative strategy of the form c_i^k for each clause.

If there exists a stable social law then it can not leave the agents with strategies of the form d_i, and must leave each agent with exactly one strategy for each clause (since otherwise a deviation to some d_j would become rational, or the agents may execute "inconsistent strategies"); these strategies need to be consistent (with respect to their corresponding assignments); hence, by combining the allowed strategies (i.e., their corresponding truth assignments) into a satisfying assignment, the other direction follows as well.

∎

Proof of Theorem 5.3 (sketch):

W.l.o.g let agent 1 be the logarithmically bounded agent. We can efficiently enumerate the set of possible restrictions on its strategies since there are only polynomially many such possibilities. For each such restriction r, let us denote the set of non-prohibited strategies by $S_1(r)$. Given $S_1(r)$ we can gather the set of strategies of the other agent (i.e., agent 2) which guarantee a payoff greater than or equals to t for agent 2 and exclude from them the ones that are dominated by other strategies of that agent (2). Let us denote this set of strategies by $S_2(r)$. Now, if there are strategies in $S_1(r)$ that are better than other strategies in $S_1(r)$ or if there exists a strategy in $S_1(r)$ which does not guarantee a payoff of t (given the previously generated set of strategies for agent 2) then we should move and try a new restriction r' on the strategies of agent 1. If that's not the case then we need to check whether there is a strategy for one of the agents which is not included in $S_1(r)$ and $S_2(r)$ respectively, and may yield a better payoff for the respective agent than what is guaranteed under $S_1(r)$ and $S_2(r)$. If there is such a deviation then we should try another r' (if exists) and otherwise we should stop (an appropriate law has been found).

The above procedure exhausts in a systematic manner all possible stable social laws since each possible restriction on the behavior of agent 1 is checked, and for each such restriction the most general restriction on the second agent's behavior which still may be possible is generated. Checking stability of a given set of restrictions is polynomial, and the above enumeration procedure is polynomial as well.

∎

Proof of Theorem 6.3 (sketch):

Assume that V' satisfying the above properties exists; one can easily check that by prohibiting all strategies in $V - V'$ we get a stable social law. The efficiency is guaranteed by the requirement from G_1, and the fact that no deviation is rational is guaranteed by the requirement from G_3. The fact that no allowed action can be ignored is guaranteed by the requirement from G_2.

If there exists a stable social law then a payoff greater than or equals to t should be guaranteed regardless of the (allowed) actions selected; this implies that the nodes associated with the allowed actions constitute a clique in G_1. Similarly, since no deviation is rational these nodes should correspond to a dominating set in

G_3. In addition, since there is no reason to consider behaviors which are inferior to others in a stable social law we get that no node which corresponds to an allowed strategy would be isolated in G_2.

∎

Proof of Theorem 6.5 (sketch):

If there is a sink in the graph then we can choose the corresponding strategy as a convention (i.e., both agents will be required to play only the corresponding strategy). Otherwise, if the graph is 2-colorable then we can color the graph by *red* and *blue* and prohibit all (and only) red strategies (for both agents). Clearly, two blue strategies will yield the desired payoff since the graph is 2-colorable. No deviation to red strategy is rational since the graph has no sinks and neighbors of a red strategy should be blue (i.e., a deviation may result in a negative payoff).

∎

References

Bond, A. H., and Gasser, L. 1988. *Readings in Distributed Artificial Intelligence.* Ablex Publishing Corporation.

Boutilier, C. 1994. Toward a Logic for Qualitative Decision Theory. In *Proc. of the 4th International Conference on Principles of Knowledge Representation and Reasoning*, 75–86.

Brafman, R., and Tennenholtz, M. 1996. On the Foundations of Qualitative Decision Theory. In *The Proceedings of AAAI-96 (to appear).*

Briggs, W., and Cook, D. 1995. Flexible Social Laws. In *Proc. 14th International Joint Conference on Artificial Intelligence*, 688–693.

Cohen, P., and Levesque, H. 1991. Teamwork. *Nous* 25(4).

Darwiche, A., and Goldszmidt, M. 1994. On the relation between kappa calculus and probabilistic reasoning. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI '94)*, 145–153.

Davis, R., and Smith, R. G. 1983. Negotiation as a metaphor for distributed problem solving. *Artificial Intelligence* 20(1):63–109.

Demazeau, Y., and Muller, J. 1990. *Decentralized AI.* North Holland/Elsevier.

Doyle, J. 1983. A Society of Mind: Multiple Perspectives, Reasoned Assumptions, and Virtual Copies. In *Proc. 8th International Joint Conference on Artificial Intelligence*, 309–314.

Dubois, D., and Prade, H. 1995. Possibility Theory as a Basis for Qualitative Decision Theory. In *Proc. 14th International Joint Conference on Artificial Intelligence*, 1924–1930.

Durfee, E. H.; Lee, J.; and Gmytrasiewicz, P. 1993. Overeager Reciprocal Rationality and Mixed Strategy Equilibiria. In *Proc. of AAAI-93*, 225–230.

Durfee, E. H.; Lesser, V. R.; and Corkill, D. D. 1987. Coherent Cooperation Among Communicating Problem Solvers. *IEEE Transactions on Computers* 36:1275–1291.

Etzioni, O.; Hanks, S.; Weld, D.; Draper, D.; Lesh, N.; and Williamson, M. 1992. An Approach to Planning with Incomplete Information. In *Proceedings of the 3rd Conference on Principles of Knowledge Representation and Reasoning*, 115–125.

Fox, M. S. 1981. An organizational view of distributed systems. *IEEE Trans. Sys., Man., Cyber.* 11:70–80.

Fudenberg, D., and Tirole, J. 1991. *Game Theory.* MIT Press.

Garey, M., and Johnson, D. 1979. *Computers and Intractability - A Guide to the Theory of NP-completeness.* W.H. Freeman and Company.

Gasser, L. 1993. Social knowledge and social action: Heterogeneity in practice. In *Proc. 13th International Joint Conference on Artificial Intelligence*, 751–757.

Genesereth, M., and Nourbakhsh, I. R. 1993. Time Saving Tips for Problem Solving with Incomplete Information. In *Proc. of AAAI-93.*

Greenberg, J. 1990. *The Theory of Social Situations; An Alternative Game-Theoretic Approach.* Cambridge University Press.

Ishida, T.; m. Yokoo; and Gasser, L. 1990. An Organizational Approach to Adaptive Production Systems. In *Proc. of AAAI-90*, 52–58.

Jennings, N. 1995. Controlling Coopeartive Problem Solving in Industrial Multi-Agent Systems Using Joint Intentions. *Artificial Intelligence* 74(2).

Kraus, S., and Wilkenfeld, J. 1991. The Function of Time in Cooperative Negotiations. In *Proc. of AAAI-91*, 179–184.

Luce, R. D., and Raiffa, H. 1957. *Games and Decisions- Introduction and Critical Survey.* John Wiley and Sons.

Malone, T. W. 1987. Modeling Coordination in Organizations and Markets. *Management Science* 33(10):1317–1332.

Minsky, N. 1991. The imposition of protocols over open distributed systems. *IEEE Transactions on Software Engineering* 17(2):183–195.

Moses, Y., and Tennenholtz, M. 1990. Artificial Social Systems Part I: Basic Principles. Technical Report CS90-12, Weizmann Institute.

Moses, Y., and Tennenholtz, M. 1995. Artificial Social Systems. *Computers and Artificial Intelligence* 14(6):533–562.

Owen, G. 1982. *Game Theory (2nd Ed.)*. Academic Press.

Peot, M. A., and Smith, D. 1992. Conditional Nonlinear Planning. In *Proceedings of the 1st International Conference on AI Planning Systems*, 189–197.

Rosenschein, J. S., and Genesereth, M. R. 1985. Deals Among Rational Agents. In *Proc. 9th International Joint Conference on Artificial Intelligence*, 91–99.

Safra, S., and Tennenholtz, M. 1994. On Planning while Learning. *Journal of Artificial Intelligence Research* 2:111–129.

Sandholm, T., and Lesser, V. 1995. Equilibrium Analysis of the Possibilities of Unenforced Exchange in Multiagent Syustems. In *Proc. 14th International Joint Conference on Artificial Intelligence*, 694–701.

Shoham, Y., and Tennenholtz, M. 1992. On Traffic Laws for Mobile Robots. Proc. of the 1st Conference on AI planning systems (AIPS-92).

Shoham, Y., and Tennenholtz, M. 1995. Social Laws for Artificial Agent Societies: Off-line Design. *Artificial Intelligence* 73.

Simon, H. A. 1981. *The Sciences of the Artificial*. The MIT Press.

Tan, S., and Pearl, J. 1994. Specification and Evaluation of Preferences under Uncertainty. In *Proc. of the 4th International Conference on Principles of Knowledge Representation and Reasoning*, 530–539.

Warren, D. H. D. 1976. Generating Conditional Plans and Programs. In *Proceedings of the Summer Conference on AI and Simulation of Behavior, Edinburgh.*

Wellman, M. P. 1993. A market-oriented programming environment and its application to distributed multicommodity flow problems. *Journal of Artificial Intelligence Research* 1:1–23.

Zlotkin, G., and Rosenschein, J. S. 1993. A Domain Theory for Task Oriented Negotiation. In *Proc. 13th International Joint Conference on Artificial Intelligence*, 416–422.