Compliance Games

Piotr Kaźmierczak

Dept. of Computing, Mathematics and Physics
Bergen University College, Norway
phk@hib.no

Abstract. In this paper we analyze *compliance games*, which are games induced by agent-labeled Kripke structures, goal formulas in the language of CTL and behavioral constraints. In compliance games, players are rewarded for achieving their goals while complying to social laws, and punished for non-compliance. Design of these games is an attempt at incentivizing agents to be compliant. We analyze the core and properties of compliance games, and study the connection between underlying logical framework and their properties.

1 Introduction

Normative systems or *social laws* are a framework for coordinating agents in multi-agent systems initially proposed by Shoham and Tennenholtz in [12,13]. The idea has been extensively studied in the multi-agent systems literature since. While in Shoham and Tennenholtz's seminal papers the framework consisted of synchronous transition systems with first order logic language for goals, in further work other semantic structures and goal languages were used. In a series of papers, Ågotnes et al. [2,3,1,5,4] presented social laws implemented on agent-labeled Kripke structures with Computation Tree Logic (CTL) as a language for goals, while Van der Hoek et al. [14] used Alternating-time Transition Systems with Alternating-time Temporal Logic (ATL) [6], and a similar framework was used in [11,8]. Each of these approaches uses the same idea, namely that we impose *restrictions* on agents' behavior,[1] and check which goals (expressed by our language of choice) are satisfied when agents comply with these restrictions.

A number of interesting decision problems are usually studied in the social laws literature, such as *compliance sufficiency* (given a structure, a set of constraints, and a goal, which coalition's compliance to the constraints is sufficient in order for the goal to be achieved?), *k-robustness* (how many agents can deviate from complying to the normative system and still not break goal satisfiability?), *feasibility* (is it feasible for the agents to satisfy their goals while complying with the restrictions?), or social law *synthesis* (can we synthesize a set of restrictions such that when complied with they guarantee goal satisfaction?).

However, the key problem in social laws is how to assure that agents comply with a given social law. Our approach here is to make compliance the rational choice for our agents. While in principle similar to the games presented by

[1] Thus social laws are sometimes also called "behavioral constraints."

Ågotnes et al. in [2] where agents had preferences over goals and normal form games were induced based on the *utility* of laws, we employ cooperative games to incentivize agents. The mechanism is simple: agents are rewarded for achieving goals while complying with laws, and punished (by means of null payoffs) for non-compliance. Formally, compliance games are induced by well-known agent-labeled Kripke structures, goal formulas expressed in the language of CTL and social laws understood as black-listed transitions of the Kripke structure. Our main contribution here is the representation theorem for compliance games and the analysis of stability (the core), which is a particularly problematic concept in this formal setting.

The paper is structured as follows. In Section 2 we provide the necessary formal background, Section 3 presents main definition of compliance games together with analysis of their properties, and in Section 4 we discuss stability of said games. We conclude and discuss future work in Section 5.

2 Technical background

We begin by concisely presenting all the formal background for our work. This paper brings temporal logic, cooperative game theory and social laws together, thus we will present a rather concise introduction to all the necessary technicalities.

2.1 Kripke structures and CTL

We start by defining agent-labelled Kripke structures, in the same way as defined by Ågotnes et al. in [3]:

Definition 2.1 (Agent-labelled Kripke Structure). *An* agent-labeled Kripke structure *(henceforth referred to simply as* Kripke structure*) K is a tuple $\langle S, S^0, R, V, \Phi, \mathcal{A}, \alpha\rangle$ where:*

- *S is the non-empty, finite set of* states *and S^0 is the initial state,*
- *$R \subseteq S \times S$ is the serial ($\forall s \exists t\ (s,t) \in R$) relation between elements of S that captures* transitions *between states,*
- *Φ is a non-empty, finite set of propositional symbols,*
- *$V : S \to 2^{\Phi}$ is a labeling function which assigns propositions to states in which they are satisfied,*
- *$\mathcal{A}$ is a non-empty finite set of agents, and*
- *$\alpha : R \to A$ is a function that labels edges with agents.*[2]

A *path* π over a relation R is an infinite sequence of states $s_0, s_1, s_2, \ldots$ such that $\forall u \in \mathbb{N} : (s_u, s_{u+1}) \in R$. $\pi[0]$ denotes the first element of the sequence, $\pi[1]$ the second, and so on. An *s-path* is a path π such that $\pi[0] = s$. $\Pi_R(s)$ is the set of s-paths over R, and we write $\Pi(s)$, if R is clear from the context.

[2] While formally not necessary, throughout the paper we assume that an agent has to "own" at least one transition.

Objectives are specified using the language of *Computation Tree Logic* (CTL), a popular branching-time temporal logic. We use an adequate fragment of the language defined by the following grammar:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathsf{E}\bigcirc \varphi \mid \mathsf{E}(\varphi \mathcal{U} \varphi) \mid \mathsf{A}(\varphi \mathcal{U} \varphi)$$

where p is a propositional symbol. The standard derived propositional connectives are used, in addition to standard derived CTL connectives such as $\mathsf{A}\bigcirc \varphi$ for $\neg\mathsf{E}\bigcirc \neg\varphi$ (see [9] for details). We distinguish two fragments of the language defined above – a *universal* L^u (with a typical element u) and an *existential* L^e (with a typical element e) one:

$$u ::= \top \mid \bot \mid p \mid \neg p \mid u \vee u \mid u \wedge u \mid \mathsf{A}\bigcirc u \mid \mathsf{A}\Box u \mid \mathsf{A}(u \mathcal{U} u)$$
$$e ::= \top \mid \bot \mid p \mid \neg p \mid e \vee e \mid e \wedge e \mid \mathsf{E}\bigcirc e \mid \mathsf{E}\Box e \mid \mathsf{E}(e \mathcal{U} e)$$

Say that we are given two Kripke structures: $K_1 = \langle S, S^0, R_1, V, \Phi, A, \alpha\rangle$ and $K_2 = \langle S, S^0, R_2, V, \Phi, A, \alpha\rangle$. We say that K_1 is a subsystem of K_2 and K_2 is a supersystem of K_1 (denoted $K_1 \sqsubseteq K_2$) if and only if $R_1 \subseteq R_2$. This yields the following observation which we will later use to prove some properties of our games.

Theorem 2.1 ([14]). *If $K_1 \sqsubseteq K_2$ and $s \in S$, then:*

$$\begin{array}{lcl} \forall e \in L^e : K_1, s \models e & \Rightarrow & K_2, s \models e; \text{ and} \\ \forall u \in L^u : K_2, s \models u & \Rightarrow & K_1, s \models u. \end{array}$$

Satisfaction of a formula φ in a state s of a structure K, $K, s \models \varphi$, is defined as follows:

$K, s \models \top$;
$K, s \models p$ iff $p \in V(s)$;
$K, s \models \neg\varphi$ iff not $K, s \models \varphi$;
$K, s \models \varphi \vee \psi$ iff $K, s \models \varphi$ or $K, s \models \psi$;
$K, s \models \mathsf{E}\bigcirc \varphi$ iff $\exists \pi \in \Pi(s) : K, \pi[1] \models \varphi$;
$K, s \models \mathsf{E}(\varphi \mathcal{U} \psi)$ iff $\exists \pi \in \Pi(s), \exists i \in \mathbb{N}$, s.t. $K, \pi[i] \models \psi$ and $\forall j, (0 \leq j < i) : K, \pi[j] \models \varphi$;
$K, s \models \mathsf{A}(\varphi \mathcal{U} \psi)$ iff $\forall \pi \in \Pi(s), \exists i \in \mathbb{N}$, s.t. $K, \pi[i] \models \psi$ and $\forall j, (0 \leq j < i) : K, \pi[j] \models \varphi$.

2.2 Social laws

A *social law* $\eta \subseteq R$ is a set of black-listed ("illegal") transitions, such that $R \setminus \eta$ remains serial. The set of all social laws over R is denoted as $N(R)$. We say that $K \dagger \eta$ is a structure with a social law η *implemented* on it, i.e.

for $K = \langle S, R, \Phi, V, A, \alpha \rangle$ and η, $K \dagger \eta = K'$ iff $K' = \langle S, R', \Phi, V, A, \alpha' \rangle$ with $R' = R \setminus \eta$ and:

$$\alpha'(s, s') = \begin{cases} \alpha(s, s') & \text{if } (s, s') \in R' \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Also, $\eta \restriction C = \{(s, s') : (s, s') \in \eta \ \&\ \alpha(s, s') \in C\}$ for any $C \subseteq A$ – that is to account for agents that do not necessarily comply with the social law (i.e. we can consider a situation in which only those edges that are "owned" by members of C are blacklisted).

2.3 Cooperative games

We now introduce some concepts from cooperative game theory. Again, definitions provided here, albeit complete, are necessarily terse. For a more detailed explanation of the concepts introduced below, see [7].

Definition 2.2. *A transferable utility* cooperative game *(sometimes also called a* coalitional*, or* characteristic function *game) is a tuple $G = \langle N, \nu \rangle$, where N is a non-empty set of* players*, and $\nu : 2^N \to \mathbb{R}$ is a* characteristic function *of the game which assigns a* value *to each coalition $C \subseteq N$ of players.*

We say a cooperative game $G = \langle N, \nu \rangle$ is *monotone* (or *increasing*) if $\nu(C) \leq \nu(D)$ whenever $C \subseteq D$ for $C, D \subseteq N$. A cooperative game is *simple* when each of the coalitions of players either wins or loses the game, in other words, when the characteristic function's signature is $\nu : 2^N \to \{0, 1\}$. Finally, we say that player i is a *veto player* in game G if $\nu(C) = 0$ for any $C \subseteq N \setminus \{i\}$.

3 Compliance Games

We now consider cooperative games in which agents are rewarded for satisfying formulas and punished for violating laws. We evaluate agents' actions based on how many of their respective *goal* formulas they are able to satisfy. Thus we say that, given a Kripke structure K, there is a set of goals:

$$\gamma_i = \{\varphi_1, \ldots, \varphi_m\}$$

associated with each agent $i \in A$ of K, where φ_j is a *goal formula* expressed in the language of CTL. We say that a Kripke structure K, a set of goals γ_i for each agent $i \in A$ of K and a social law η over R of K constitute a *social system* $\mathcal{S} = \langle K, \gamma_1, \ldots, \gamma_n, \eta \rangle$.

Below we introduce the definition of the game. The idea behind it is that the value of a coalition C is the amount of goals it can achieve under restrictions minus the amount of goals achievable in a Kripke structure (without restrictions). This number can be negative, and the rationale behind such design of games is that behavior of agents would indicate to the system designer whether the laws he designed are optimal or not (i.e., if an agent can satisfy more of his goals while not complying than when complying then perhaps either the goals or the laws need to be adjusted).

Definition 3.1 (Compliance Game). *A social system $\mathcal{S}$ induces a cooperative game $G_{\mathcal{S}} = \langle \mathcal{A}, \nu_{\mathcal{S}} \rangle$, where $\mathcal{A}$ is a set of agents in K of $\mathcal{S}$, and*

$$\nu_{\mathcal{S}}(C) = \begin{cases} 1 & \text{if } \left(\sum_{\varphi \in \gamma_C} \left| \{\varphi : K \dagger (\eta \restriction C) \models \varphi\} \right| - \sum_{\varphi \in \gamma_C} \left| \{\varphi : K \models \varphi\} \right| \right) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $C \subseteq \mathcal{A}$, and $\gamma_C = \bigcup_{i \in C} \gamma_i$.

We now analyze some properties of compliance games. First we observe that the characteristic function of said games is not always monotone.

Lemma 3.1. *Compliance games are not always monotone.*

Proof. The lemma above can be proved with a simple counter example shown in Figure 1. As seen in the example, adding agents to a winning coalition can "break" the satisfiability of their goals. □

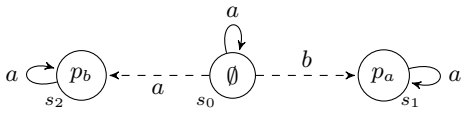

Fig. 1. A Kripke structure which induces a non-monotone compliance game, illustrating Lemma 3.1. Here, $\gamma_a = \{\mathsf{E}\bigcirc p_a\}$, $\gamma_b = \{\mathsf{E}\bigcirc p_b\}$, $K \models \gamma_a$, $K \models \gamma_b$, and $K \dagger (\eta \restriction \{a\}) \models \gamma_a$, $K \dagger (\eta \restriction \{b\}) \models \gamma_b$, but $K \dagger (\eta \restriction \{a, b\}) \not\models \gamma_a \vee \gamma_b$.

The fact that compliance games are not always monotone is a negative result from a point of view of algorithmic game theory, because we cannot take computational advantage of the monotonicity property of the characteristic function. In fact we present a Theorem below which states something even stronger:

Theorem 3.1. *Given an arbitrary function $\nu : 2^{\mathcal{A}} \to \{0, 1\}$, there is always a social system $\mathcal{S} = \langle K, \gamma_1, \ldots, \gamma_n, \eta \rangle$ which induces a cooperative game $G_{\mathcal{S}} = \langle A, \nu_{\mathcal{S}} \rangle$, where A is a set of agents in K of $\mathcal{S}$ and $\nu = \nu_{\mathcal{S}}$.*

Proof. We prove the theorem by providing a recipe for creating a social system $\mathcal{S} = \langle K, \gamma_1, \ldots, \gamma_n, \eta \rangle$ in which K is constructed of elements which "isolate" winning conditions for each winning coalition.

We construct $\mathcal{S}$ in the following way. Each agent is given the same goal: $\mathsf{E}\Diamond\mathsf{A}\Box p$, thus γ_i is a singleton set which contains one formula. We then construct the Kripke structure K starting from the initial state which is not labeled by any proposition and has a transition labeled by an arbitrary agent and not blacklisted by η (a "bridge") leading to another state s_C (again, not labeled by any proposition). The s_C state is the beginning of a construction which assures that coalition C wins. We then construct a sequence of states not labeled by any

propositions with transitions labeled by all agents from $\mathcal{A} \setminus C$, one per agent, all of which are included in the social law η (this assures that the superset of C does not win along this path). Once we are done, we add a state labeled by all the goals of members of C and a reflexive loop labeled by an arbitrary agent. This state becomes the satisfied goal once members of C comply to η. Next, in order to assure subsets of C lose the game, we add a state and a transition per member of C leading to a state in which the negation of all the goals is satisfied, labeling transitions with respective agents and adding them to η – this way only if all members of C comply the transitions will be blacklisted and the goal satisfied. This whole construction from s_0 assures winning conditions for coalition C, and we can construct a separate such construction for each winning C'. Any C' that does not have a construction of this kind is a losing coalition, thus we can model *any* set of outcomes of the game. The construction is presented in Figure 2. □

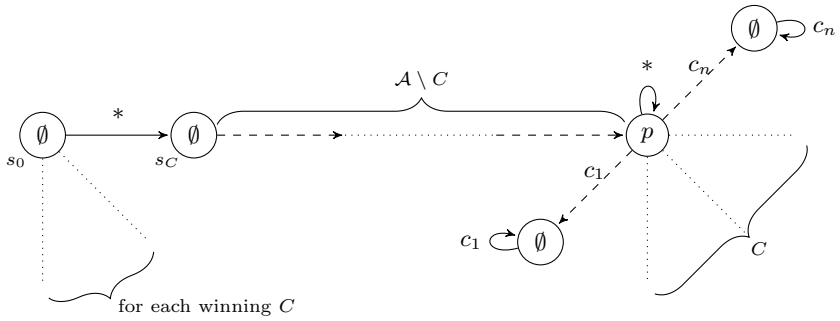

Fig. 2. Construction for the proof of Theorem 3.1. Dashed lines represent transitions in η, the $*$ symbol stands for an arbitrary agent.

The representation result presented in this section is more general, since it can easily be adapted to similar simple games. In [1], authors present social systems $S = \langle K, \varphi, \eta \rangle$ and induce simple cooperative games of the form

$$\nu_S(C) = \begin{cases} 1 & \text{if } K \dagger (\eta \restriction C) \models \varphi \\ 0 & \text{otherwise,} \end{cases}$$

in order to study power indices for agents. Since the game defined above is very similar to our compliance game, we immediately obtain the following result:

Corollary 3.1. *Given an arbitrary function $\nu : 2^{\mathcal{A}} \to \{0, 1\}$, there is always a social system $S = \langle K, \varphi, \eta \rangle$ which induces a cooperative game $G_S = \langle \mathcal{A}, \nu_S \rangle$ as defined in [1], where $\mathcal{A}$ is a set of agents in K of S.*

Since monotonicity is a desirable property for most coalitional games, we are interested in identifying subclasses of our games that are actually monotone.

Ågotnes et al. [3] identify social systems with universal goals as a subclass of monotone versions of their games. We can do the same for our compliance games.

Proposition 3.1. *Compliance games are monotone for* $\varphi \in L^u$.

Proof. Let $\mathcal{S} = \langle K, \eta, \gamma_1, \ldots, \gamma_n \rangle$ be a social system, and let $C \subseteq C' \subseteq \mathcal{A}$ be coalitions in K. Then from Theorem 2.1 we know that if $K \dagger (\eta \upharpoonright C) \models u$ then $K \dagger (\eta \upharpoonright C') \models u$. Thus if an agent's universal goal becomes satisfied by his compliance (removal of edges), another agent's compliance cannot break satisfiability of that goal. For the same reason if there is a universal goal satisfied in a Kripke structure, imposing a social law will not change its satisfiability. □

4 Stability – the core

We concentrate on the most popular stability related solution concept, which is *the core*. In order to define it precisely, we need a few extra definitions.

We say that an *imputation* is a vector $(p_1, \ldots, p_n)$ with $p_i \in \mathbb{Q}$ which is such a division of gains of a grand coalition N that $\sum_{i=1}^{n} p_i = \nu(N)$. We say that p_i is a *payoff* for player i and denote the payoff for coalition C as $p(C) = \sum_{i \in C} p_i$.

An imputation satisfies *individual rationality* when for all players $i \in C$, we have $p_i \geq \nu(\{i\})$. A coalition B *blocks* a payoff vector $(p_1, \ldots, p_n)$ when $p(B) < \nu(B)$ – members of B can abandon the original coalition, with each member getting p_i and there is still some amount of $\nu(B)$ to be shared amongst players. The coalition is thus unstable when a blocking payoff vector is chosen.

Definition 4.1. *The* core *of a cooperative game is a set of such imputations which are not blocked by any coalition. Formally, for an imputation* $(p_1, \ldots, p_n)$*, and any coalition* C*,* $p(C) \geq \nu(C)$*.*

Intuitively, the core characterizes such a set of outcomes where no player has an incentive to abandon the coalition structure. Many games have an empty core, however simple games usually have a non-empty core, due to the following well-known result:

Lemma 4.1 ([10]). *A core of a simple cooperative game is non-empty iff there is at least one veto player in the game. And if there are veto players, any imputation that distributes payoffs only to them is in the core of the game.*

Thus checking non-emptiness of the core for a simple game boils down to finding whether there are veto players.

Before we attempt to identify outcomes in the core of compliance games, we need to discuss in detail what "stability" of said games actually means. Recall that in Definition 3.1 the value for a coalition is defined as the difference between how many goals a coalition can achieve while complying to a social law and the amount of goals achievable when not complying. There is an important detail in the semantics of $K \dagger (\eta \upharpoonright C)$ expression, though, that makes an interpretation of stability-capturing solution concepts somewhat problematic.

As mentioned in the paragraphs above, the core characterizes a set of stable outcomes, where stable means players have no incentives to "deviate" and pursue goals on their own. What does "deviating" mean in the context of compliance games? It means complying, but with other players. However, the expression $K \dagger (\eta \restriction C)$ means that players in coalition C comply with η, but at the same time *no other players comply*. This means that if a proper subset of coalition C intends to abandon its coalition and form a new coalition C', the rationale behind its agents' behavior is to form a coalition in which they themselves comply, and at the same time they somehow force all other players to *not* comply. This aspect of compliance games is highly unusual and makes the interpretation of the core non-standard. The way we motivate and explain such design of compliance games is that we are interested in said games from a system designer's point of view. Reasoning this way, we may interpret a particular coalition structure of a compliance game as a set of *hypothetical scenarios*: each winning coalition is such a scenario, and in this scenario the coalition in question assumes no one but its members comply.

Proposition 4.1. *Compliance games can have empty cores.*

Proof. From Lemma 4.1 we know that the only situation when a compliance game has an empty core is when there are no veto players present in the game. From Theorem 3.1 we know that we can create a social system that induces a compliance game with a characteristic function that has an arbitrary output, so e.g. for a social system with two players a and b we can construct $\nu(\{a\}) = \nu(\{b\}) = 1$, but $\nu(\{a, b\}) = 0$ and $\nu(\mathcal{A}) = 0$. □

In light of this negative result we would be interested in a relationship between the shape of elements of $\mathcal{S}$ and the non-emptiness of the core of the compliance game it induces, or more narrowly, we are interested in the relationship between the structure of $\mathcal{S}$ and the existence of veto players.

The first likely suspects to yield games with non-empty cores are social systems with only universal goal formulas, since these games will be monotone, as shown in Proposition 3.1. In monotone games it is easy to see that player i is a veto player iff $\nu(\mathcal{A} \setminus \{i\}) = 0$, thus any game with at least one winning and one losing coalition will have at least one veto player.

Proposition 4.2. *Let $\mathcal{S}$ be a social system containing only universal goals. If $\mathcal{S}$ induces a compliance game where there is at least one winning and one losing coalition, then this game's core is non-empty.*

Proof. Follows directly from the statement of the problem and Proposition 3.1. □

5 Conclusions and future work

In this paper we presented a new, game theoretic approach to incentivizing agents' compliance to social laws. We analyzed properties of compliance games,

studied their stability and relation between their properties and the underlying logical framework that induces them. For future work, we plan to first investigate complexity of decision problems and also try designing compliance games based on other classes of cooperative games. We think that the general idea of using cooperative game theory for modeling compliance to social laws is a naturally attractive and to a great degree unexplored idea, and we wish to pursue it further in the future.

Acknowledgments The author would like to thank Pål Grønås Drange, Hannah Hansen, Truls Pedersen, and reviewers of EUMAS 2014 for helpful comments.

References

1. T. Ågotnes, W. van der Hoek, M. Tennenholtz, and M. Wooldridge. Power in normative systems. In *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 145–152, 2009.
2. T. Ågotnes, W. van der Hoek, and M. Wooldridge. Normative system games. In *Proc. of the 6th Int. Joint Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2007)*, pages 1–8, 2007.
3. T. Ågotnes, W. van der Hoek, and M. Wooldridge. Robust normative systems and a logic of norm compliance. *Logic Journal of the IGPL*, 18(1):4–30, 2009.
4. T. Ågotnes, W. van der Hoek, and M. Wooldridge. Conservative Social Laws. In *Proc. of the 20th European Conf. on Artificial Intelligence (ECAI 2012)*, pages 49–54, 2012.
5. T. Ågotnes and M. Wooldridge. Optimal Social Laws. In *Proc. of the 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, pages 667–674, 2010.
6. R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time Temporal Logic. *Journal of the ACM*, 49(5):672–713, 2002.
7. G. Chalkiadakis, E. Elkind, and M. Wooldridge. *Computational Aspects of Cooperative Game Theory*. Number 16 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012.
8. S. Dyrkolbotn and P. Kaźmierczak. Playing with Norms: Tractability of Normative Systems for Homogeneous Game Structures. In *Proc. of the 13th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, pages 125–132, 2014.
9. E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science Volume B: Formal Models and Semantics*. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1990.
10. M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.
11. T. Pedersen, S. Dyrkolbotn, and P. Kaźmierczak. Big, but not unruly: Tractable norms for anonymous game structures. arXiv:1405.6899, 2013.
12. Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies. In *Proc. of the 10th National Conf. on Artificial intelligence (AAAI 1992)*, pages 276–281, 1992.
13. Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73:231–252, 1995.
14. W. van der Hoek, M. Roberts, and M. Wooldridge. Social laws in alternating time: effectiveness, feasibility, and synthesis. *Synthese*, 156(1):1–19, Sept. 2006.