Game-Theoretic Foundations for Norms

Guido Boella
Dipartimento di Informatica
Università di Torino-Italy
E-mail: guido@di.unito.it

Leendert van der Torre
Department of Computer Science
University of Luxembourg
E-mail: leendert@vandertorre.com

Abstract

In this paper we study game-theoretic foundations for norms. We assume that a norm is a mechanism to obtain desired multi-agent system behavior, and must therefore under normal or typical circumstances be fulfilled by a range of agent types, such as norm internalizing agents, respectful agents fulfilling norms if possible, and selfish agents obeying norms only due to the associated sanctions.

1 Introduction

The relation between game theory and norms has received some attention. E.g., in a widely discussed example of the so-called centipede game, there is a pile of thousand pennies, and two agents can in turn either take one or two pennies. If an agent takes one then the other agent takes turn, if it takes two then the game ends. A backward induction argument implies that it is rational only to take two at the first turn. Norms and trust have been discussed to analyze this behavior, see [6] for a discussion.

Our approach to study this relation it to use game-theory for the foundations of normative systems. In artificial social systems or normative multi-agent systems, a social law or norm is a mechanism to achieve desired system behavior. Since in an open system it cannot be assumed that agents obey the norm, there has to be a control system motivating agents to obey the norm, by monitoring and sanctioning behaviors. Moreover, the system should not sanction without reason, as for example Caligula or Nero did in the ancient Roman times, as the norms would loose their force to motivate agents. Various ways to motivate agents including norms have been studied in economics, for example using the game-theoretic machinery to study the rationality of norms.

The research question of this paper is how to give game-theoretic foundations to norms such as obligations, permissions and counts-as conditionals. Most of our study is focussed on obligations, and therefore on incentives like sanctions and rewards.

We first consider the so-called partially controlled multi-agent system (PCMAS) approach of Brafman and Tennenholtz [3], one of the classical game-theoretical studies of social laws in so-called artificial social systems developed by Tennenholtz and colleagues, because incentives like sanctions and rewards play a central role in this theory. So-called controllable agents – agents controlled by the system programmer – enforce social behavior by punishing and rewarding agents, and thus can be seen as representatives of the normative system. For example, consider an iterative prisoner dilemma. A controlled agent can be programmed such that it defects when it happens to encounter an agent which has defected in a previous round.

The PCMAS model thus distinguishes between two kinds of agent interaction in the game theory, namely between two normal (so-called uncontrollable) agents, and between a normal and a controllable agent. We show in this paper that this makes it a very useful model to give game-theoretic foundations to norms. Whereas classical game theory is only concerned with interaction among normal agents, it is the interaction among normal and controllable agents which we use in our game theoretic foundations.

The PCMAS approach not only clarifies the design of punishments, but it also illustrates the iterative and multi-agent character of social laws. However, there are also drawbacks of the model, such that it cannot be used to give a completely satisfactory game-theoretic foundation for norms. We would like to express that a norm can be used for various kinds of agents, such as norm internalizing agents, respectful agents that attempt to evade norm violations, and selfish agents that obey norms only due to the associated sanctions. Therefore, as classical game theory is too abstract to satisfactorily distinguish among agent types, we consider also cognitive agents and qualitative game theory.

The layout of this paper is as follows. We introduce a qualitative game-theory based on a logic for mental attitudes and cognitive agent theory, which we use to give game-theoretic foundations of obligations and permissions.

2 Qualitative games among cognitive agents

In Boella and Lesmo's game-theoretic approach to norms [2], a rational definition of sanction-based obligations is given using classical game theory by representing the normative system as an agent. They model the normative system as a set of controlled agents, as in the PCMAS model, but they do not necessarily assume that they are controlled by the system programmer. We use a model of cognitive agents that is able to distinguish among norm internalizing agents, respectful agents that attempt to evade norm violations, and selfish agents that obey norms only due to the associated sanctions.

We have to be brief on technical details, and refer the reader to other work for the details. The important issue here is to give the flavor of cognitive agent theory, where the maximization of expected utility is replaced by maximization of achieved goals.

2.1 Input/output logic for mental attitudes

Makinson and van der Torre [9] define the proof theory of input/output logic as follows.

Definition 1 *Let L be a propositional language, let the norms in G be pairs of L $\{(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)\}$, read as 'if input α_1, then output β_1', etc., and consider the following proof rules strengthening of the input (SI), conjunction for the output (AND), weakening of the output (WO), disjunction of the input (OR), and cumulative transitivity (CT) and Identity (Id) defined as follows:*

$$\frac{(\alpha,\gamma)}{(\alpha\wedge\beta,\gamma)}SI \qquad \frac{(\alpha,\beta),(\alpha,\gamma)}{(\alpha,\beta\wedge\gamma)}AND \qquad \frac{(\alpha,\beta\wedge\gamma)}{(\alpha,\beta)}WO$$

$$\frac{(\alpha,\gamma),(\beta,\gamma)}{(\alpha\vee\beta,\gamma)}OR \qquad \frac{(\alpha,\beta),(\alpha\wedge\beta,\gamma)}{(\alpha,\gamma)}CT \qquad \frac{}{(\alpha,\alpha)}Id$$

The following four output operators are defined as closure operators on the set G using the rules above.

out_1: *SI+AND+WO*	*(simple-minded output)*
out_2: *SI+AND+WO+OR*	*(basic output)*
out_3: *SI+AND+WO+CT*	*(reusable output)*
out_4: *SI+AND+WO+OR+CT*	*(basic reusable output)*

Moreover, the following four throughput operators are defined as closure operators on the set G.

out_i^+: out_i*+Id (throughput)*

We write $out(G)$ for any of these output operations, and $out^+(G)$ for any of these throughput operations.

Semantics of input/output logics have been given for $out(G)$ in a classical Tarskian style (a model is a pair of sets of propositional valuations, with additional constraints) and for $out(G, A) = \{x \mid a \subseteq A, (a, x) \in out(G)\}$ in a more operational style. Moreover, extensions of input/output logics have been developed for contrary-to-duty reasoning [10] and for reasoning about weak and various kinds of strong permission [11].

The following definition extends constraints with a priority relation among norms, to resolve conflicts. Moreover, it introduces undercutter rules $a \rightarrow b$, which mean that if input a, then the output does *not* contain x. They are used to model permissions as exceptions to obligations.

Definition 2 *Let L be a propositional language, let G and H be two sets of pairs of L, let $\geq$: $2^{G \cup H} \times 2^{G \cup H}$ be a transitive and reflexive relation on subsets of these pairs, and let out be an output operation.*

- *A pair $\langle G', H' \rangle$ is consistent in a if $out(G', a)$ is consistent, and for each $(b, x) \in H'$, if $b \in out(G', a)$ then $x \notin out(G', a)$.*
- *$maxfamily(G, H, \geq, a)$ is the set of pairs $\langle G' \subseteq G, H' \subseteq H \rangle$ that:*
  1. *are consistent in a, and*
  2. *if $\langle G'' \subseteq G, H'' \subseteq H \rangle$ is consistent in a, $G' \subseteq G''$, and $H' \subseteq H''$, then $G'' = G'$ and $H'' = H'$;*

  In other words, it is maximal with respect to set inclusion among the consistent pairs;
- *$preffamily(G, H, \geq, a)$ is the set of pairs $\langle G', H' \rangle$ that:*
  1. *are in $maxfamily(G, H, \geq, a)$, and*
  2. *if $\langle G'' \subseteq G, H'' \subseteq H \rangle$ is in $maxfamily(G, H, \geq, a)$ and $G'' \cup H'' \geq G' \cup H'$, then $G'' = G'$ and $H'' = H'$;*

  In other words, it is maximal with respect to $\geq$;
- *$out_{\cap}(G, H, \geq, a) = \cap out(preffamily(G, H, \geq, a), \geq, a)$; In other words, using only the rules which occur in all elements of preffamily.*

2.2 Beliefs, goals, decisions, decision rule

To represent that agents are autonomous decision makers, we associate a set of decision variables with each agent. Decisions or actions are based on controllability from control theory or discrete event systems (not to be confused with controllable agents!). Moreover, each agent has four sets of rules, besides beliefs and goals also undercutters for beliefs and goals. Finally, each agent has a priority relation among these rules.

Definition 3 *Let L be a propositional logic based on the set of propositions X. A multi-agent system is a tuple $\langle A, B, G, C, H, AD, MD \geq \rangle$ where:*

- *the agents A, beliefs, B, goals G, belief undercutters C, and goal undercutters H are five disjoint sets;*
- *the agent description $AD : A \rightarrow 2^{B \cup G \cup C \cup H \cup X}$ is a function from agents to its beliefs, goals, undercutters, and decision variables;*
- *the mental description $MD : B \cup C \cup G \cup H \rightarrow L \times L$ is a function from the mental attitudes to input/output rules;*
- *the priority relation $\geq : A \rightarrow (2^{B \cup C} \times 2^{B \cup C}) \cup (2^{G \cup H} \times 2^{G \cup H})$ is a binary relation on sets of beliefs and goals.*

The qualitative decision rule is based on maximizing achieved goals, or minimizing unachieved goals.

Definition 4 *Given a multi-agent system, and let out^+ be a throughput operation for beliefs, and out an output operation for goals.*

- *A decision profile $\delta : (\cup_{a \in A} AD(a) \cap X) \rightarrow \{0,1\}$ is a function from the decision variables to truth values; We represent δ by a logical formula;*
- *The expected effects of decision profile δ for agent a, $E(\delta, a)$, are $out^+_\cap(AD(a) \cap B, AD(a) \cap C, \geq, \delta)$;*
- *The unachieved goals according to agent a, $U(\delta, a)$, are $\cap(G \setminus \{G' | \langle G', H' \rangle \in$ preffamily$(G, H, \geq, E(\delta, a))\})$*
- *δ is preferred to δ' iff $U(\delta, a) \subset U(\delta', a)$*

2.3 Agent types

The qualitative game theory works analogous to the classical game theory, where the maximization of expected utility is replaced by a minimization of unachieved goals. This more detailed model allows us to distinguish among various agent types. In this paper, we consider three agent types.

First, we consider norm internalizing agents. These uncontrollable agents incorporate some of the goals of the controllable agents. They thus behave like controllable agents, if it is in their power.

Second, respectful agents try to fulfill obligations when they can do so. We model this by making the violation conditions explicit in the controllable agents. They do not sanction directly, but they first determine whether observed behavior is a violation, and then associate a sanction with a violation. Respectful agents obey the norm, if they can, regardless of the sanction.

Third, selfish agents care only about the sanctions imposed by the controllable agents. They behave as traditional agents in economic theory.

3 Six clauses for obligation

For obligation and prohibition, we need at least six clauses. The first clause ensures that “respectful” agents internalizing the goals of the normative system will fulfill their obligation under typical circumstances, the second and third clause do so for “respectful” agents not internalizing the norm, and the other clauses do so for “selfish” types of agents. The first clause says that the obligation is in the desires and in the goals of a normative system b (“your wish is my command”). The second and third clause can be read as “the absence of x is considered as a violation”. The association of obligations with violations is inspired by Anderson’s reduction of deontic logic to alethic modal logic [1]. The third clause says that the normative system desires that there are no violations. The fourth and fifth clause relate violations to sanctions and assume that normative system b is motivated to apply sanctions only as long as there is a violation; otherwise the norm would have no effect. Finally, for the same reason, we assume in the last clause that the agent does not like the sanction.

Definition 5 (Obligation) *Let MAS be a multi-agent system, and $G_a = AD(a) \cap G$, etc. Agent $a \in A$ is obliged in MAS to decide to do x with sanction s if Y by controllable agent b, written as $MAS \models O_{a,b}(x, s \mid Y)$, if and only if:*

1. $Y \to x \in out(G_b)$: *if controllable agent b believes Y, then it desires and has as a goal that x.*

2. $Y \wedge x \to V_a(\neg x) \in out(G_b)$: *if controllable b believes Y and $\neg x$, then it has the goal $V_a(\neg x)$: to recognize $\neg x$ as a violation by agent a.*

3. $\top \to \neg V_a(\neg x) \in G_b$: *controllable agent b desires that there are no violations.*

4. $Y \wedge V_a(\neg x)\} \to s \in out(G_b$: *if controllable agent b believes Y and decides $V_a(\neg x)$, then it desires that it sanctions agent a with s.*

5. $Y \wedge \neg s \in out(G_b)$: *if controllable agent b believes Y, then it desires not to sanction, $\neg s$. The controllable agent only sanctions in case of violation.*

6. $Y \to \neg s \in out(G_a)$: *if agent a believes Y, then it desires $\neg s$, which expresses that it does not like to be sanctioned.*

4 Two clauses for permission

We do not define permissions as the absence of obligation, so-called negative permission, but as exceptions to obligations, a kind of positive permission. For a discussion on the issues involved in modeling permission, see [11]. Permission is simpler than obligation, since permissions cannot lead to violations and sanctions.

Here we distinguish between permission and entitlement or right. It is only due to entitlement that knowledge providers may be sanctioned when they do not permit a user to access documents, but the user itself cannot be a violator and be sanctioned due to its permissions to access a document. which distinguishes between users that are only permitted to access knowledge, and users that are also entitled to it in the sense that knowledge providers are obliged to permit them access. Games can be played to show that the clauses of permission are necessary, again for norm internalizing agents and other types of agents respectively.

Definition 6 (Permission) *Let MAS be a multi-agent system. Agent $a \in A$ is permitted to decide to do x if Y in MAS by controllable agent b, written as $MAS \models P_{a,b}(x \mid Y)$, if and only if:*

1. *$Y \rightarrow x \in out(H_b)$: if controllable agent b believes Y, then it does not have a goal that x.*
2. *$Y \cup \{x\} \rightarrow \neg V_a(x) \in out(G_b)$: if controllable agent b believes Y and x, then it does not want to count x as a violation.*

5 Summary

We show how the distinction between controllable and uncontrollable agents can be used to give game-theoretic foundations for norms. First, we discuss how the PCMAS model can be used to give such foundations for obligations. The main drawback of this model is that it is not clear how to distinguish among agent types. Another drawback is that it does not explain how to deal with other kinds of norms, such as permissions.

Second we discuss a straightforward extension of the PCMAS model using goal based reasoners instead of utility maximizers. We show also how game-theoretic considerations can be used to model permissions, based on a discussion by Bulygin. The use of qualitative game theory makes a bridge to deontic logic, which formalized logical relations among obligations and permissions.

A third kind of norms are constitutive norms known as counts-as conditionals, which define institutional facts in a normative system, for example which pieces of paper count as money, and which relations count as marriages. The game theoretic foundations of this kind of norms is subject of further research.

References

[1] A.R. Anderson. The logic of norms. *Logic et analyse*, 2, 1958.

[2] G. Boella and L. Lesmo. A game theoretic approach to norms. *Cognitive Science Quarterly*, 2(3-4):492–512, 2002.

[3] Ronen I. Brafman and Moshe Tennenholtz. On partially controlled multi-agent systems. *J. Artif. Intell. Res. (JAIR)*, 4:477–507, 1996.

[4] R Conte, C Castelfranchi, and F Dignum. autnomous norm acceptance. 1998.

[5] Gneezy and Rustichini. A fine is a price. *Journal of Legal Studies*, XXIX(1):1–18, 2000.

[6] M. Hollis. *Trust within Reason*. Cambridge University Press, 1998.

[7] A. Jones and J. Carmo. Deontic logic and contrary-to-duties. In D. Gabbay, editor, *Handbook of Philosophical Logic*, pages 203–279. Kluwer, 2002.

[8] Levitt and Dubner. *Freakonomics*. William Morrow, New York, 2005.

[9] D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.

[10] D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic*, 30(2):155–185, 2001.

[11] D. Makinson and L. van der Torre. Permissions from an input/output perspective. *Journal of Philosophical Logic*, 32 (4):391–416, 2003.

[12] Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73 (1-2):231 – 252, 1995.

[13] Y. Shoham and M. Tennenholtz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94 (1-2):139 – 166, 1997.

[14] M. Tennenholtz. On stable social laws and qualitative equilibria. *Artificial Intelligence*, 102 (1):1–20, 1998.