International Journal of Cooperative Information Systems

NORM NEGOTIATION IN MULTIAGENT SYSTEMS

GUIDO BOELLA
Dipartimento di Informatica - Università di Torino - Italy. E-mail: guido@di.unito.it

LEENDERT VAN DER TORRE
University of Luxembourg. E-mail: leendert@vandertorre.com

Received (Day Month Year)
Revised (Day Month Year)

Normative multiagent systems provide agents with abilities to autonomously devise societies and organizations coordinating their behavior via social norms and laws. In this paper we study how agents negotiate new social norms and when they accept them. We introduce a negotiation model based on what we call the social delegation cycle, which explains the negotiation of new social norms from agent desires in three steps. First individual agents or their representatives negotiate social goals, then a social goal is negotiated in a social norm, and finally the social norm is accepted by the agents when it leads to fulfilment of the desires the cycle started with. We characterize the allowed proposals during social goal negotiation as mergers of the individual agent desires, and we characterize the allowed proposals during norm negotiation as both joint plans to achieve the social goal (obligations associated with the norm) and the associated sanctions or rewards (a control system associated with the norm). The norm is accepted when the norm is stable in the sense that agents will act according to the norm, and effective in the sense that fulfilment of the norm leads to achievement of the agents' desires. We also compare norm negotiation with contract negotiation and negotiation of the distribution of obligations.

1. Introduction

Social norms and laws can be used to guide the emergent behavior of multiagent systems. Moreover, these norms guiding the emergent behavior of the system can also emerge themselves, leading to a challenge for models of normative multiagent systems. To deal with the apparent recursion in the definition of the emergence of norms, we need a model making it precise when and how norms can emerge. Since agents may have conflicting goals with respect to the norms that emerge, they try to negotiate amongst each other which norm will emerge. Moreover, we may say that a norm has emerged when it has been accepted by the agents, as has been studied, for example, by Conte *et al.*[14]

The negotiation and acceptance of norms has to obey various principles. For example, to accept a norm, an agent has to recognize it as a norm, the norm must contribute to the goals or desires of the agent, and it must be obeyed by the other agents. Consequently, the challenge for a model of norm negotiation is how to explain what it means, for example, to recognize or to obey a norm. Moreover, there are additional challenges, for example how

new norms interact with existing ones.

In this paper we study norm emergence and acceptance using a model of norm negotiation. It is based on a distinction between social goal negotiation and the negotiation of the obligations with their control system. Roughly, the social goals are the potential benefits of the new norm for the agents, and the obligations are the potential costs of the new norm for the agents in the sense that agents risk being sanctioned if they do not respect the norm. Moreover, in particular when representatives of the agents negotiate the social goals and norms, the agents still have to accept the negotiated norms.[14] The norm is accepted when the norm is stable in the sense that agents will act according to the norm, and effective in the sense that fulfilment of the norm leads to achievement of the agents' desires – i.e., when the benefits outweigh the costs. Summarizing, norm negotiation can be described by the so-called social delegation cycle visualized in Figure 1.

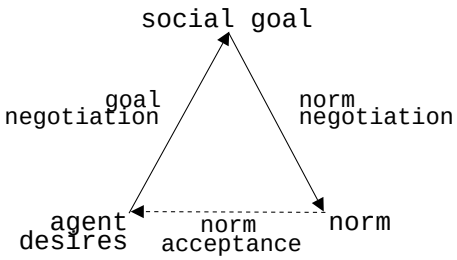

Fig. 1. Δ: the social delegation cycle.

The conceptual model we use to study and formalize the social delegation cycle is based on a formal characterization of normative multiagent systems we have developed in several other papers.[7,8] This characterization combines a game-theoretic approach with cognitive agents, rule based systems and input/output logics. The advantage given by this formal system is that it combines a qualitative theory of agent decision-making with a detailed ontology of normative systems. The decision making is used to explain how agents negotiate goals and norms, and when they accept norms.

Within the model of normative multiagent systems, we have to choose a protocol for negotiation. Since we are not interested in the intricacies of negotiation itself, but in the relation among goal negotiation, norm negotiation and acceptance, we choose a relatively simple abstract turn-taking model which we can use for both negotiation steps. The central issue in such an abstract negotiation model is the definition of an allowed proposal for a social goal or for a norm, for which we define additional conditions.

The layout of this paper is as follows. In Section 2 we give an introduction to normative multiagent systems, and in Section 3 we discuss an abstract model of the social delegation cycle. In Section 4 we define the conceptual model in which we study and formalize the social delegation cycle, and in Section 5 we define our version of the negotiation protocol and the logic of rules. In Section 6 we formalize goal negotiation, in Section 7 we formalize norm negotiation, and in Section 8 we formalize the acceptance relation.

2. Normative multiagent systems

Jones and Carmo[21] define a normative system as "Sets of agents whose interactions are norm-governed; the norms prescribe how the agents ideally should and should not behave. [...] Importantly, the norms allow for the possibility that actual behavior may at times deviate from the id eal, i.e., that violations of obligations, or of agents' rights, may occur." Since the agents' control over the norms is not explicit here, we use the following definition.

> A normative multiagent system is a multiagent system together with normative systems in which agents can decide whether to follow the explicitly represented norms, and the normative systems specify how and in which extent the agents can modify the norms.[9]

Note that this definition makes no presumptions about the internal architecture of an agent or of the way norms find their expression in agent's behavior.

Since norms are explicitly represented, according to our definition of a normative multiagent system, the question should be raised how norms are represented. Norms can be interpreted as a special kind of constraint, and represented depending on the domain in which they occur. However, the representation of norms by domain dependent constraints runs into the question what happens when norms are violated. Not all agents behave according to the norm, and the system has to deal with it. In other words, norms are not hard constraints, but soft constraints. For example, the system may sanction violations or reward good behavior. Thus, the normative system has to monitor the behavior of agents and enforce the sanctions. Also, when norms are represented as domain dependent constraints, the question will be raised how to represent permissive norms, and how they relate to obligations. Whereas obligations and prohibitions can be represented as constraints, this does not seem to hold for permissions. For example, how to represent the permission to access a resource under an access control system? Finally, when norms are represented as domain dependent constraints, the question can be raised how norms evolve.

We therefore believe that norms should be represented as a domain independent theory. For example, deontic logic[40,38,37,25,26,27] studies logical relations among obligations and permissions, and more in particular violations and contrary-to-duty obligations, permissions and their relation to obligations, and the dynamics of obligations over time. Therefore, insights from deontic logic can be used to represent and reason with norms. Deontic logic also offers representations of norms as rules or conditionals. However, there are several aspects of norms which are not covered by constraints nor by deontic logic, such as the relation between the cognitive abilities of agents and the global properties of norms.

Conte *et al.*[14] say that normative multiagent systems research focuses on two different sets of problems. On the one hand, they claim that legal theory and deontic logic supply a theory for of norm-governed interaction of autonomous agents while at the same time lacking a model that integrates the different social and normative concepts of this theory. On the other hand, they claim that three other problems are of interest in multiagent systems research on norms: how agents can acquire norms, how agents can violate norms, and how an agent can be autonomous. For artificial agents, norms can be designed as in legal human

systems, forced upon, for example when joining an institution, or they can emerge from the agents making them norm autonomous.[39] Agent decision making in normative systems and the relation between desires and obligations has been studied in agent architectures,[10] which thus explain how norms and obligations influence agent behavior.

An important question is where norms come from. Norms are not necessarily created by a single legislator, they can also emerge spontaneously, or be negotiated among the agents. In electronic commerce research, for example, cognitive foundations of social norms and contracts are studied.[7] Protocols and social mechanisms are now being developed to support such creations of norms in multiagent systems. When norms are created, the question can be raised how they are enforced. For example, when a contract is violated, the violator may have to pay a penalty. But then there has to be a monitoring and sanctioning system, for example police agents in an electronic institution. Such protocols or roles in a multiagent system are part of the construction of social reality, and Searle[32] has argued that such social realities are constructed by constitutive norms. This again raises the question how to represent such constitutive or counts-as norms, and how they are related to regulative norms like obligations and permissions.[8]

Not only the relation between norms and agents must be studied, but also the relation between norms and other social and legal concepts. How do norms structure organizations? How do norms coordinate groups and societies? How about the contract frames in which contracts live? How about the relation between legal courts? Though in some normative multiagent systems there is only a single normative system, there can also be several of them, raising the question how normative systems interact. For example, in a virtual community of resource providers each provider may have its own normative system, which raises the question how one system can authorize access in another system, or how global policies can be defined to regulate these local policies.[8]

Summarizing, normative multiagent systems study general and domain independent properties of norms. It builds on results obtained in deontic logic, the logic of obligations and permissions, for the representation of norms as rules, the application of such rules, contrary-to-duty reasoning and the relation to permissions. However, it goes beyond logical relations among obligations and permissions by explaining the relation among social norms and obligations, relating regulative norms to constitutive norms, explaining the evolution of normative systems, and much more.

Within multiagent systems, norms and social laws are used to design social mechanisms, and in normative systems, agent theory is used to analyze and simulate social phenomena. Various examples can be found in a forthcoming double special issue of *Computational and Mathematical Organizational Theory* on normative multiagent systems.[9] The models most closely related to the one presented in this paper can be found in work on artificial social systems by Tennenholtz and colleagues, and the work on social theory by Castelfranchi, Conte and colleagues, which we discuss in the following section. The detailed model of normative systems including a sanction-based control system together with the game-theoretic aspects like acceptance and backward induction distinguishes our approach from numerous other approaches of social norms and laws in multiagent systems.[16,17,18,19,24,33]

3. Social delegation cycle

To motivate our model, we start with an abstract description of the social delegation cycle using 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[33] introduce social laws in a setting without utilities, and they define *rational* social laws as social laws that improve a social game variable.[34] We follow Tennenholtz' presentation of artificial social systems for stable social laws.[36]

The desires of the agents are represented by their utilities. 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 observes also that the extension to the multi-agent case is straightforward.[36]

Definition 3.1. 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 $U_1 : S \times T \to I\!R$ and $U_2 : S \times T \to I\!R$ are real-valued utility functions for agents 1 and 2, respectively.

The social goal is represented by a minimal value for the social game variable. Tennenholtz[36] uses as game variable the maximin value. 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 3.2. 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.

The social norm is represented by a social law, which has 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 3.3. 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 $\overline{S} \subseteq S$, and of $\overline{T} \subseteq T$. The social law is *useful* if the following holds: there exists $s \in \overline{S}$ such that $U_1(s, \overline{T}) \geq e$, and there exists $t \in \overline{T}$ such that $U_2(\overline{S}, t) \geq e$. A (useful) convention is a (useful) social law where $|\overline{S}| = |\overline{T}| = 1$.

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 3.4. 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 $\overline{S}$ and T to $\overline{T}$, and satisfies the following: there is no $s' \in S - \overline{S}$ which satisfies $U_1(s', \overline{T}) > \max_{s \in \overline{S}} U_1(s, \overline{T})$, and there is no $t' \in T - \overline{T}$ which satisfies $U_2(\overline{S}, t') > \max_{t \in \overline{T}} U_2(\overline{S}, t)$.

The three steps of the social delegation cycle in this classical game-theoretic setting can be identified as follows.

Goal negotiation implies that the efficiency parameter is higher than the utility the agents expect without the norm, for example represented by the Nash equilibria of the game.

Norm negotiation implies that the social law is useful (with respect to the efficiency parameter).

Acceptance relation implies that the social law is quasi-stable.

An important class of examples from game theory are coordination problems, like the choice between driving on either the right hand or the left hand side of the road. Such problems are characterized by the fact that it does not matter which option we choose, as long as everyone chooses the same option. Another classic example from game theory is the prisoner's dilemma, which is characterized by the fact that optimal situations can be reached only when we introduce sanctions on behaviors (such as defecting). Both of these examples can be modeled in artificial social systems. Without norms the agents expect accidents and defects, and therefore in goal negotiation the efficiency parameter is set such that there are no accidents, and no defections. The norm negotiation implies that the agent should all drive on the right hand side of the street and they should not defect (otherwise they are sanctioned), and finally the agents accept the norm, because they know that another agent will obey it (if all other agents do so too).

However, there are also drawbacks of this game-theoretic model of the social delegation cycle. Due to the uniform description of agents in the game-theoretic model, it is less clear how to distinguish among kinds of agents. For example, the unique utility aspiration level does not distinguish the powers of agents to negotiate a better deal for themselves than for the other agents. Moreover, the formalization of the social delegation cycle does neither give a clue how the efficiency parameter is negotiated, nor how the social law is negotiated. For example, the desires of the agents as well as other mental attitudes may play a role in this negotiation. Finally, an additional drawback is that the three ingredients of the model (agent desires, social goals, and social laws) are formalized in three different ways. These drawbacks also seem to hold when a normative system to encode enforceable social laws is added to the artificial social system model.[6]

To make a more realistic representation of the above notions, some models use a more detailed cognitive agent theory, in which agents are motivated by desires and goals, and norms are explicitly represented inside agents. For example, Conte *et al.*[14] discuss under which conditions new norms can be acquired by cognitive agents by recognizing them and accepting that they become goals. Conte *et al.*'s first condition for norm recognition concerns the evaluation of the source. They say that "if the norm is not based upon a recognized norm, the entity y that has issued the norm is evaluated. If y is perceived to

be entitled to issue norms (it is a normative authority), then [the obligation issued by y and addressed to the set of agents X to obtain q, written as] $O_{yX}(q)$, can be accepted as a norm." Their second relevant issue in the recognition of a norm is the evaluation of the motives for which the norm has been issued. "$O_{yX}(q)$ is issued for y's personal/private interest, rather than for the interest y is held to protect: if x believes that y's prescription is only due to some private desire, etc., x will not take it as a norm. x might ignore what the norm is for, what its utility is for the group or its institutions, but may expect that the norm is aimed at having a positive influence for the group; at least, it is necessary that x does not have the opposite belief, that is, that the norm is not aimed to be 'good for' the group at large, but only for y. ... y is entitled only to deliver prescriptions and permissions that are aimed at the general rather than at its own private interest."

Conte *et al.* claim that for the acceptance of norms as goals to be considered in the decision process requires, instead, the following stronger condition.

> "A norm-autonomous agent accepts a norm q, only if it sees accepting q as a way of achieving (one of) its own further goal(s)."

This condition is formalized as follows, where $OBT_X(q)$ means that the agents X obtain q. It "states that x forms a normative goal $OBT_X(q)$ (i.e. accepts the norm q) if x believes that the norm exists (for agents in set X) and that fulfilling the norm (i.e. $OBT_X(q)$) is instrumental to one of its own goals."

$$BEL_x(O_{yX}(q) \wedge INSTR(OBT_X(q), p) \wedge GOAL_x(p|r)) \supset \\ N\text{-}GOAL_x(OBT_X(q)|GOAL_x(p|r) \wedge r)$$

Conte *et al.* emphasize that they "do not require q to be instrumental for the goal p, but rather $OBT_X(q)$. With $OBT_X(q)$ in this context we mean the fulfilment of the norm q by all members of X. The difference is that in this case we only try to fulfil the norm, because it is a norm. We could also have the much stronger case in which we believe that the norm itself is to the benefit of our goal p. This is somehow 'internalising' the norm and making it our own goal. This would formally be described by:

$$BEL_x(O_{yX}(q) \wedge INSTR(q, p) \wedge GOAL_x(p|r)) \supset C\text{-}GOAL_x(q|GOAL_x(p|r) \wedge r) \quad "$$

Conte *et al.* observe that the achievement of the norm can be instrumental to some goal of the agent in different ways:

(1) Arbitrated instrumentality: "x sees accepting q as a means to obtain a non-natural consequent reward. In particular, avoidance of punishment, social praise, etc.," where they distinguish among avoidance of external or internal punishment, and achievement of external or internal reward.
(2) Natural instrumentality, where they distinguish self-interested from value-oriented.

Finally, Conte *et al.* observe that a norm must have general recognition. "Without a general recognition, a social norm is not a norm. (At the legal level it is sufficient that the authority is recognised as authority and that the norm is recognised as correctly issued by

it). Autonomous agents subject to norms are in fact autonomous norm creators. They create norms through their evaluation and recognition, through their compliance, and through their interpersonal issuing, monitoring and judging."

From this cognitive setting we learn the following for the social delegation cycle:

Goal negotiation: To be a norm, a norm must not be issued in the interest of the authority only. Thus for an agent to recognize a norm as such, the norm must be aimed to be good for the group at large.

Norm negotiation: The creation of a norm must lead to a norm which achieves some agent own's goals: for example, the goal not to be considered as a violator and sanctioned, or to be rewarded.

Acceptance relation: To be a norm the norm must be recognized, accepted and complied with in general. Thus, the norm is created also by means of the agents' behavior.

Given the cognitive perspective, Conte *et al.*'s model allows a deeper understanding of the acquisition of norms. It leaves, however, several issues informal. In particular, what does it mean that a norm is "good for" the group at large"? What does it means that it is recognized in general? In general, how can a game-theoretic analysis be added to a cognitive one to understand how norms are created?

In this paper we answer these questions in a cognitive agent framework. In our model agent desires, social goals and social laws are all represented by qualitative rules, formalized using input/output logics.[25] To formalize and characterize the negotiation of social goals and of obligations together with their sanctions we have to formalize the negotiation protocol including the allowed proposals: proposals which do not meet general interests cannot be made. Our model builds on several existing formal theories to define the acceptable proposals:

Goal negotiation is based on proposals where the social goal is obtained using merging operators, which have been proposed as generalizations of belief revision operators inspired by social choice theory.

Norm negotiation is based on proposals that create for each social goal a set of obligations (or revisions of existing obligations) together with sanctions or rewards.

Acceptance relation accepts or rejects a norm based on the stability criterion of artificial social systems and the general acceptance of cognitive theories, together with a cost-benefit analysis of the agents that the costs of having to comply to the norm are outweighed by the benefit of the social goal of the norm.

As a running example, we consider three sets of agents, who can work together in various ways. They can make a coalition to each perform a task, or they can distribute five tasks among them and obtain an even more desirable social goal. There can be free riders, so they need sanctions to ensure that the norm is accepted.

We define in the following sections the conceptual framework based on rule based systems, the generic negotiation protocol used both for social goal negotiation and norm negotiation, and the logic of rules based on input/output logics.

4. Conceptual model of normative multiagent systems

The conceptual model used in this paper is visualized in Figure 2, in which we distinguish the multiagent system (straight lines) and additions for the normative system (dotted lines). Following the usual conventions of for example class diagrams in the unified modelling language (UML), □ is a concept or set, — and → are relations or associations among concepts, and —▷ is the "is-a" or subset relation. The logical structure of the associations is detailed in the definitions below.

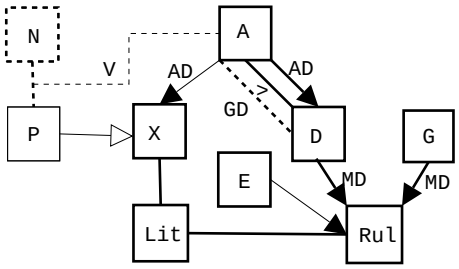

Fig. 2. Conceptual model of normative multiagent system.

The model consists of a set of agents (A) described (AD) by a set of boolean variables (X), including *decision variables* it can perform, and desires (D) guiding its decision making. Desire rules can be conflicting, and the way the agent resolves its conflicts is described by a priority relation ($\geq$) that expresses its agent characteristics[10]. The priority relation is defined on the powerset of the motivations such that a wide range of characteristics can be described, including social agents that take the desires of other agents or social goals into account. The priority relation contains at least the subset-relation which expresses a kind of independence between the motivations. Variables which are not decision variables are called parameters (P).

Definition 4.1. (AS) An *agent set* is a tuple $\langle A, X, D, AD, \geq \rangle$, where:

- the agents A, variables X and agent desires D are three finite disjoint sets.
- an agent description $AD : A \rightarrow 2^{X \cup D}$ is a complete function that maps each agent to sets of variables (its decision variables) and desires, but that does not necessarily assign each variable to at least one agent. For each agent $a \in A$, we write X_a for $X \cap AD(a)$, and D_a for $D \cap AD(a)$. We write parameters $P = X \setminus \cup_{a \in A} X_a$.
- a priority relation $\geq$: $A \rightarrow 2^D \times 2^D$ is a function from agents to a transitive and reflexive relation on the powerset of the desires containing at least the subset relation. We write $\geq_a$ for $\geq (a)$.

The motivational state of the society is composed of its social goals (G), and background knowledge is formalized by a set of effect rules (E). Desires and social goals are abstract concepts which are described by – though conceptually not identified with – rules

(Rul) built from literals (Lit). They are therefore not represented by propositional formulas, as in some other approaches to agency.[13,29] Agents may share decision variables, or desires, though this expressive power is not used in this paper.

Definition 4.2. (MAS) A *multiagent system* is a tuple $\langle AS, G, E, MD\rangle$, where $AS = \langle A, X, D, AD, \geq\rangle$ is an agent set, and:

- the set of goals G is a set disjoint from A, X, and D. The motivations $M = D \cup G$ are defined as the union of the desires and social goals.
- the set of literals built from X, written as $Lit(X)$, is $X \cup \{\neg x \mid x \in X\}$, and the set of rules built from X, written as $Rul(X) = 2^{Lit(X)} \times Lit(X)$, is the set of pairs of a set of literals built from X and a literal built from X, written as $\{l_1, \ldots, l_n\} \rightarrow l$. We also write $l_1 \wedge \ldots \wedge l_n \rightarrow l$ and when $n = 0$ we write $\top \rightarrow l$. Moreover, for $x \in X$ we write $\sim x$ for $\neg x$ and $\sim(\neg x)$ for x.
- the set of effects $E \subseteq Rul(X)$ is a set of rules built from X.
- the motivational description $MD : M \rightarrow Rul(X)$ is a complete function from the sets of desires and goals to the set of rules built from X. For a set of motivations $S \subseteq M$, we write $MD(S) = \{MD(s) \mid s \in S\}$.

To describe the normative system, we introduce a set of norms (N) and a norm description that associates violations with variables (V).

Definition 4.3. (*NMAS*) A normative multiagent system is a tuple $\langle MAS, N, V\rangle$, where $MAS = \langle AS, G, E, MD\rangle$ is a multiagent system, $AS = \langle A, X, D, AD, \geq\rangle$ is an agent set, and moreover:

- the set of norms N is a set disjoint from A, X, D, and G.
- the norm description $V : N \times A \rightarrow P$ is a complete function that maps each pair of a norm and an agent to the parameters, where $V(n, a) \in P$ represents that a violation by agent a of the norm n has been recognized.

We define sanction- and reward-based obligations in the normative multiagent system using an extension of Anderson's well-known reduction[2,28]: violations and sanctions are the consequences of not fulfilling a norm. It covers a kind of ought-to-do and a kind of ought-to-be obligations.

Definition 4.4. (Obligation) Let $NMAS = \langle MAS, N, V\rangle$, $MAS = \langle AS, G, E, MD\rangle$ and $AS = \langle A, X, D, AD, \geq\rangle$. We say that:

- x is obligatory for agent a in $NMAS$ iff $\exists n \in N$ with $\sim x \rightarrow V(n, a) \in E$,
- s is a sanction for agent a in $NMAS$ iff $\exists n \in N$ with $V(n, a) \rightarrow s \in E$, and
- r is a reward for agent a in $NMAS$ iff $\exists n \in N$ with $\neg V(n, a) \rightarrow r \in E$.

More sophisticated notions within this kind of framework are developed elsewhere.[7] In this paper we now turn to the negotiation protocol.

5. Generic negotiation protocol and logic of rules

A negotiation protocol is described by a sequence of negotiation actions which either lead to success or failure. In this paper we only consider protocols in which the agents propose a so-called deal, and when an agent has made such a proposal, then the other agents can either accept or reject it. Moreover, they can also end the negotiation process without any result. Such negotiation protocols can be represented by a finite state machine, but we do not consider such representations in this paper. For turn-taking, we assume that the agents involved are ordered ($\succeq$).

Definition 5.1. (Protocol) A negotiation protocol is a tuple $NP = \langle Ag, deals, actions, valid, finished, broken, \succeq \rangle$ where:

- the agents Ag, $deals$ and $actions$ are three disjoint sets, such that $actions = \{propose(a, d), accept(a, d), reject(a, d) \mid a \in Ag, d \in deals\} \cup \{breakit(a) \mid a \in Ag\}$.
- $valid$, $finished$, $broken$ are three sets of finite sequences of $actions$.
- $\succeq \subseteq Ag \times Ag$ is a total order on Ag,

We instantiate this generic protocol for negotiations in normative multiagent systems. We assume that a sequence of actions (a history) is valid when each agent does an action respecting its turn in the ordering of agents. Then, after each proposal, the other agents have to accept or reject this proposal, again respecting the ordering of agents, until they all accept it or one of them rejects it. When it is an agent's turn to make a proposal, it can also end the negotiation by breaking it. The history is $finished$ when all agents have accepted the last deal, and $broken$ when the last agent has ended the negotiations.

Definition 5.2. (NMAS protocol) Let $NMAS$ be a normative multiagent system $\langle MAS, N, V \rangle$ together with $MAS = \langle AS, G, E, MD \rangle$ and $AS = \langle A, X, D, AD, \geq \rangle$. A negotiation protocol $NP = \langle Ag, deals, actions, valid, finished, broken \rangle, \succeq \rangle$ for $NMAS$ satisfies the following constraints:

- the set of agents $Ag \subseteq A$ consists of the negotiators representing the agents A;
- A sequence h ends iff either all agents have accepted the last proposal ($finished(h)$) or the last agent has broken the negotiation ($broken(h)$) instead of making a new proposal.
- a history h is a sequence of actions, and $valid(h)$ holds if:
  - the $propose$ and $breakit$ actions in the sequence respect $\succeq$,
  - each $propose$ is followed by a sequence of $accept$ or $reject$ actions respecting $\succeq$ until either all agents have accepted the deal or one has rejected it,
  - there is no double occurrence of a proposal $propose(a, d)$ of a deal $d \in deals$ by any agent $a \in Ag$, and
  - the sequence h ends.

The open issue of the generic negotiation protocol is the set of deals which can be proposed. They depend on the kind of negotiation. In social goal negotiation the deals represent a social goal, and in norm negotiation the deals contain the obligations of the

agents and the associated control system based on sanctions. To define the deals, we have to be more precise about the logic of rules we use.

We use a simplified input/output logic.[25,26] A rule base is a set of rules, i.e., a set of ordered pairs $p \rightarrow q$. For each such pair, the body p is thought of as an input, representing some condition or situation, and the head q is thought of as an output, representing what the norm tells us to be desirable, obligatory or whatever in that situation. We use input/output logics since they do not necessarily satisfy the identity rule. Makinson and van der Torre write (p, q) to distinguish input/output rules from conditionals defined in other logics, to emphasize the property that input/output logic does not necessarily obey the identity rule. In this paper we do not follow this convention.

In this paper, input and output are respectively a set of literals and a literal. We use a simplified version of input/output logics, since it keeps the formal exposition simple and it is sufficient for our purposes here. In Makinson and van der Torre's input/output logics, the input and output can be arbitrary propositional formulas, not just sets of literals and literal as we do here. Consequently, in input/output logic there are additional rules for conjunction of outputs and for weakening outputs.

Definition 5.3. (Input/output logic[25]) Let a rule base B be a set of rules $\{p_1 \rightarrow q_1, \ldots, p_n \rightarrow q_n\}$, read as 'if input p_1 then output q_1', *etc.*, and consider the following proof rules strengthening of the input (SI), disjunction of the input (OR), and cumulative transitivity (CT) defined as follows:

$$\frac{p \rightarrow r}{p \wedge q \rightarrow r} SI \qquad \frac{p \wedge q \rightarrow r, p \wedge \neg q \rightarrow r}{p \rightarrow r} OR \qquad \frac{p \rightarrow q, p \wedge q \rightarrow r}{p \rightarrow r} CT$$

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

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

We write $out(B)$ for any of these output operations, we write $B \vdash_{iol} p \rightarrow q$ for $p \rightarrow q \in out(B)$, and we write $B \vdash_{iol} B'$ iff $B \vdash_{iol} p \rightarrow q$ for all $p \rightarrow q \in B'$.

The following definition of the so-called input-output constraint checks whether the derived conditional goals are consistent with the input.

Definition 5.4. (Constraints[26]) Let B be a set of rules, and C a set of literals. B is consistent with C, written as $cons(B|C)$, iff there do not exist two contradictory literals p and $\neg p$ in the set $C \cup \{l | B \vdash_{iol} C \rightarrow l\}$. We write $cons(B)$ for $cons(B|\emptyset)$.

The semantics of input/output logics, further details on its proof theory, the extension with the identity rule, alternative constraints, permissions, and examples are described elsewhere.[25,26,27] We do not consider the negotiation of permissions or constitutive norms in this paper.

6. Social goal negotiation

We characterize the allowed deals during goal negotiation as a merger or fusion of the desires of the agents, which may be seen as a particular kind of social choice process.[23] Technically, we use the merging operators for merging desires into goals in the context of beliefs,[15,23], which are generalizations of belief revision operators.[1,20,22]. We simplify these operators, because we do not use beliefs, and we make them more complex, because we extend the operators defined on propositional formulas to merge rules.

Definition 6.1. A rule base B is a set of desire rules of an agent, a rule set S of a multiagent system is a multi-set of rule bases. Two rule sets S_1 and S_2 are equivalent, written as $S_1 \leftrightarrow S_2$, iff there exists a bijection f from $S_1 = \{B_1^1, \ldots, B_1^n\}$ to $S_2 = \{B_2^1, \ldots, B_2^n\}$ such that $out(f(B_1^i)) = out(B_1^i)$ for $i = 1 \ldots n$. We write $\bigwedge S$ for the union of all rules in S, and $\sqcup$ for union with multi-sets.

A merging operator is defined by a set of postulates the result of the merger has to obey. For example, the social goal should not violate the integrity constraints (R0 in Definition 6.2.), when the integrity constraints are consistent we should always manage to extract a coherent social goal from the rule set (R1), and if possible, the social goal is simply the conjunction of the rule bases of the rule set with the integrity constraints (R2). Moreover, the principle of irrelevance of syntax says that if we change the syntax but not the meaning of the desire rules of the agents, the social goal does not change (R3).

Koniecnzy[23] introduces the fairness postulate, which ensures that when merging two rule bases, the operator cannot give full preference to one of them (R4). Pareto's conditions in Arrow's social choice theory[3] detail how mergers of sets of rules are related (R5 and R6). Finally there are conditions on the conjunction of integrity constraints (R7 and R8). See the above mentioned papers for further details and motivations. In the following definition, as well as in all following definitions, we assume that a logic of rules has been fixed.

Definition 6.2. Let $\vdash_{iol}$ be an output operation, S be a rule set, E a rule base, and ∇ an operator that assigns to each rule set S and rule base E a rule base $\nabla_E(S)$. ∇ is a rule merging operator if and only if it satisfies the following properties:

R0 If not $cons(E)$, then $\nabla_E(S) \leftrightarrow E$
R1 If $cons(E)$, then $cons(\nabla_E(S))$
R2a $\bigwedge S \vdash_{iol} \nabla_E(S)$
R2b If $cons(\bigwedge S \cup E)$, then $\nabla_E(S) \vdash_{iol} \bigwedge S$
R3 If $S_1 \leftrightarrow S_2$ and $E_1 \leftrightarrow E_2$, then $\nabla_{E_1}(S_1) \leftrightarrow \nabla_{E_2}(S_2)$
R4 If $B \vdash_{iol} E$, $B' \vdash_{iol} E$, and $cons(\nabla_E(\{B\} \sqcup \{B'\}) \cup B \cup E)$, then $cons(\nabla_E(\{B\} \sqcup \{B'\}) \cup B' \cup E)$
R5 $\nabla_E(S_1) \cup \nabla_E(S_2) \vdash_{iol} \nabla_E(S_1 \sqcup S_2)$
R6 If $cons(\nabla_E(S_1) \cup \nabla_E(S_2) \cup E)$, then $\nabla_E(S_1 \sqcup S_2) \vdash_{iol} \nabla_E(S_1) \cup \nabla_E(S_2)$
R7 If $cons(E_1 \cup E_2)$, then $\nabla_{E_1}(S) \vdash_{iol} \nabla_{E_1 \cup E_2}(S)$
R8 If $cons(\nabla_{E_1}(S) \cup E_1 \cup E_2)$, then $\nabla_{E_1 \cup E_2}(S) \vdash_{iol} \nabla_{E_1}(S)$

Additional properties and the semantics of merging operators have been defined in the

literature.[23] In this paper, we only consider social goals consisting of one rule. At first sight this may seem like a strong restriction, but more complex examples can be modeled using effect rules in E.

Definition 6.3. (Deals in goal negotiation) In the goal negotiation protocol, the type of the allowed deals is that deals are a rule base consisting of a single rule.

Let $NMAS$ be a normative multiagent system $\langle MAS, N, V\rangle$ together with $MAS = \langle AS, G, E, MD\rangle$ and $AS = \langle A, X, D, AD, \geq\rangle$. We have $d \in deals$ iff there is a rule merging operator ∇ such that $MD(d) \subseteq \nabla_E(MD(D(x)) \mid x \in A)$.

The following example illustrates that the allowed deals depend on the chosen input/output logic.

Example 6.1. Let $\vdash_{iol}$ be out_3, and consider four agents with desires represented by rule bases each consisting of a single rule $S = \{\{\top \rightarrow p\}, \{\top \rightarrow q\}, \{p \rightarrow r\}, \{q \rightarrow \neg r\}\}$. We cannot take all rules in S as the social goal, since they are not consistent. Formally, due to (R1) we cannot propose the deal $\nabla_\emptyset(S) = \bigwedge S$, because we do not have $cons(\bigwedge S)$. We can take for example either the deal $\nabla_\emptyset(S) = \{\top \rightarrow p, \top \rightarrow q, p \rightarrow r\}$, or the deal $\nabla_\emptyset(S) = \{\top \rightarrow q, p \rightarrow r, q \rightarrow \neg r\}$, since they are consistent and they satisfy the other postulates. For example, they satisfy (R4) since there are no two rule bases directly conflicting with each other. In general, checking whether a social goal base satisfies the merging postulates is non-trivial, and can be based on the semantics of merging operators.[23] If we assume that $\vdash_{iol}$ is out_1, then $cons(\bigwedge S)$, and due to (R2b) we have $\nabla_\emptyset(S) = \bigwedge S$.

Social goal negotiation is illustrated by our running example. We consider three sets of agents, who can work together in various ways. They can make a coalition to each perform a task, or they can distribute five tasks among them and obtain an even more desirable social goal. Variables in the example are *syntactic sugar*. Quantification over rules means that rules are schemata: there is a set of rules, one for each agent involved. Note that the set of agents A is finite.

Example 6.2. (Running example) Let $NMAS = \langle MAS, N, V\rangle$ together with $MAS = \langle AS, G, E, MD\rangle$ and $AS = \langle A_1 \cup A_2 \cup A_3, X, D, AD, \geq\rangle$ be a normative multiagent system with, amongst others, the following ingredients:

variables: $X = \{task_1(a), task_2(a), task_3(a) \mid a \in A_1\} \cup \ldots$
Each agent in A_1 can perform task 1,2,3, each agent in A_2 can perform task 2,3,4, and each agent in A_3 can perform task 3, 4, and 5.

effect rules: $E = \{\ldots \cup \wedge_{i=1\ldots5} task_i \rightarrow best\_state\}$
Performing task 1, 3 and 5 will lead to some good result, and all five tasks will lead to the best result.

desires: for $a \in A_1$, $MD(D_a) = \{task_4, task_5\}$,....
Each agent desires the tasks it cannot perform itself.

Social goal $MD(G) = \{\top \rightarrow best\_state\}$.
The agents negotiate the social goal to perform all five tasks for $NMAS$.

7. Norm negotiation

We formalize the allowed deals during norm negotiation as solutions of a planning problem, distinguishing between the obligations and the associated sanctions and rewards. The planning problem for the obligations is that the obligations of the agents must imply the social goal. We represent a norm n by an obligation for all agents in the multiagent system, that is, for every agent a we introduce an obligation $\sim x \rightarrow V(n,a)$. Moreover, since goals can only be in force in a context, e.g., $Y \rightarrow g$, we introduce in context Y an obligation $Y \wedge \sim x \rightarrow V(n,a)$. Roughly, the condition is that the conjunction of all obligations x imply the social goal g.

However, to determine whether the obligations imply the social goal, we have to take the existing normative system into account. We assume that the normative system only creates obligations that can be fulfilled together with the already existing obligations. Moreover, for the test that the social goal g will be achieved, we propose the following condition: if every agent fulfills its obligation, and it fulfills all its other obligations, then g is achieved. We define a global violation constant $\mathbf{V}$ as the disjunction of all indexed violation constants like $V(n,a)$, i.e., $\mathbf{V} = \bigvee_{n \in N, a \in A} V(n,a)$.

Note that the type of the proposed deals in norm negotiation is more complex than the type in goal negotiation. In goal negotiation each agent can propose a social goal, which is a rule base. In norm negotiation each agent can propose a norm, which contains obligation rules together with sanction rules (for each agent).

Definition 7.1. (Deals in norm negotiation with sanctions and rewards) In norm negotiation, the type of $deals$ is a pair of rule bases.

Let $NMAS$ be a normative multiagent system $\langle MAS, N, V \rangle$ together with $MAS = \langle AS, G, E, MD \rangle$ and $AS = \langle A, X, D, AD, \geq \rangle$, and let $Y \rightarrow g \in MD(G)$ be a social goal. We assume that the parameters contain the global violation constant $\mathbf{V} \in P$ and E contains the following set of rules:

$$\{V(n,a) \rightarrow \mathbf{V} \mid n \in N, a \in A\} \cup \{\neg \mathbf{V} \rightarrow \neg V(n,a) \mid n \in N, a \in A\}$$

We have $\langle E', E'' \rangle \in deals$ when E' is a set of obligations and E'' is a set of sanctions and rewards, defined as follows,

$$E' \subseteq \{Y \wedge x \rightarrow V(n',a) \mid a \in A, x \in Lit(X)\}$$

$$E'' \subseteq \{Y \wedge V(n',a) \rightarrow s \mid a \in A, s \in Lit(X)\} \cup \{Y \wedge \neg V(n',a) \rightarrow r \mid a \in A, r \in Lit(X)\}$$

and, moreover:

(1) The norm n' is not already part of N;
(2) E' is a set of obligations for each $a \in A$ such that $E \cup E' \vdash_{iol} \neg \mathbf{V} \wedge Y \rightarrow g$, if all norms are fulfilled, then the social goal is fulfilled;
(3) $cons(E \mid Y \wedge \neg \mathbf{V})$, it is possible that no norm is violated.
(4) E'' is a set of sanctions and rewards for each $a \in A$ such that for all such s and r we have $D_a \vdash_{iol} Y \rightarrow \neg s$ or $D \vdash_{iol} Y \rightarrow r$, sanctions are undesired and rewards are desired.

If the norm n' with obligations E' and sanctions and rewards E'' to achieve social goal $Y \rightarrow g$ in $deals$ is accepted, then it leads to the updated system $\langle MAS, N \cup \{n'\}, V \rangle$ with $MAS = \langle AS, E \cup E' \cup E'', MD \rangle$.

The following example illustrates norm negotiation in our running example. There can be free riders, so they need sanctions to ensure that the norm is accepted.

Example 7.1. (Continued) The normative multiagent system is further detailed as follows (we assume that the same social goal is negotiated):

Desires: $MD(D) = \ldots \cup \{\top \rightarrow \neg task_1(a), \top \rightarrow \neg task_2(a), \ldots \mid a \in A_1\} \cup \ldots$
Each agent desires not to perform any tasks. These desires are weaker (have lower priority in $\succeq$) than the desire that the other tasks are performed. Moreover, each agent desires not to be sanctioned, written as $s(a)$.

Effect rules
We now assume that not all agents have to contribute to a task, but only most of them. Thus there can be free riders.

E.g., a possible deal contains the obligations $E' = \{\neg task_1(a) \rightarrow V(n', a) \mid a \in A_1\}$ and sanctions $E'' = \{V(n', a) \rightarrow s(a) \mid a \in A\}$. $\neg task_1(a)$ counts as a violation of norm n' by agent a, and is sanctioned with $s(a)$.

The following example illustrates the negotiation protocol.

Example 7.2. (Continued) Consider three agents that have to negotiate a task consisting of five subtasks,with social goal $g = \top \rightarrow best\_state$, the $NMAS\ protocol$ is:
$Ag = \{a_1, a_2, a_3\}$ with $a_1 \in A_1$, $a_2 \in A_2$, $a_3 \in A_3$,
$deals$ = the sets of pairs $\langle \tau_\delta, \tau_\sigma \rangle$ where the set of obligations τ_δ belongs to the set $\Delta = \{\delta = \{task_1(a_1), task_2(a_2), task_3(a_3), task_4(a_4), task_5(a_5)\} \mid a_1, \ldots, a_5 \in Ag\}$. Moreover, τ_σ is a tuple of $|A|$ elements specifying a sanction $s_i \in X$ for every agent in A. Here is a history h, where a_1 proposes something which is not accepted, but a_2 thereafter proposes a distribution which is accepted:

$action_1 : propose(a_1, d_1 = \langle \tau_\delta, \langle s_1, s_2, s_3 \rangle \rangle)$ where
$\quad \tau_\delta = \{task_1(a_1), task_2(a_2), task_3(a_3), task_4(a_3), task_5(a_3)\}$
$action_2 : accept(a_2, d_1)$
$action_3 : reject(a_3, d_1)$
$action_4 : propose(a_2, d_2 = \langle \tau'_\delta, \langle s_1, s_2, s_3 \rangle \rangle)$ where
$\quad \tau'_\delta = \{task_1(a_1), task_2(a_2), task_3(a_2), task_4(a_3), task_5(a_3)\}$
$action_5 : accept(a_3, d_2)$
$action_6 : accept(a_1, d_2)$

We have $valid(h)$, because the order of action respects $\leq$, and we have $accepted(h)$, because the history ends with acceptance by all agents ($action_5$ and $action_6$) after a proposal ($action_4$).

8. Norm acceptance and agent decision making

An agent accepts a norm when the obligation implies some desire the cycle started with, and moreover, it believes that the other agents will fulfill their obligations. We propose the following games: agent a plays a game with arbitrary agent b and accepts the norm if agent b fulfills the norm *given that all other agents fulfill the norm*, and this fulfilment leads to fulfillment of agent a's desire the cycle started with. This implies that the fulfillment of the goal g is kind of normative equilibrium. The details and examples of the decision model and the notion of unfulfilled desires for conditional rules can be found in our work on normative multiagent systems.[7,8]

Definition 8.1. (Decision) Let $NMAS$ be a normative multiagent system $\langle MAS, N, V\rangle$ with $MAS = \langle AS, G, E, MD\rangle$ and $AS = \langle A, X, D, AD, \geq\rangle$. The optimal decision of agent $b \in A$ given a set of literals C is defined as follows.

- The set of decisions Δ is the set of subsets of $Lit(X_b)$ that do not contain a variable and its negation. A decision is complete if it contains, for each variable in X_b, either this variable or its negation.
- The unfulfilled desires of decision δ for agent $b \in A$, written as $U(\delta, b)$, are the desires whose body is part of the decision, but whose head is not.

$$\{d \in D_b \mid MD(d) = L \rightarrow l, E \vdash_{iol} C \cup \delta \rightarrow l' \text{ for } l' \in L \text{ and } E \nvdash_{iol} C \cup \delta \rightarrow l\}$$

- A decision δ is *optimal* for agent b if and only if there is no decision δ' such that $U(\delta, b) >_b U(\delta', b)$.

We use the definition of optimal decision to define the acceptance relation. We define a variant of the global violation constant $\mathbf{V}_{\sim b}$ as the disjunction of the violation constants of all agents except agent b. We assume here that the agents only consider typical cases. In reality there are always exceptions to the norm, but we do not take this into account.

Definition 8.2. (Acceptance) Let $NMAS$ be a normative multiagent system $\langle MAS, N, V\rangle$ with $MAS = \langle AS, G, E, MD\rangle$ and $AS = \langle A, X, D, AD, \geq\rangle$, and let $NMAS' = \langle MAS', N \cup \{n'\}, V\rangle$ with $MAS' = \langle AS, E \cup E' \cup E'', MD\rangle$ be the system after the creation of a norm. The parameters contain the global violation constants $\mathbf{V}_{\sim b} \in P$ and E contains the following rules:

$$\{V(n, x) \rightarrow \mathbf{V}_{\sim b} \mid n \in N, x \in A \setminus \{b\}\} \cup \{\neg\mathbf{V}_{\sim b} \rightarrow \neg V(n, x) \mid n \in N, x \in A \setminus \{b\}\}$$

An agent $a \in A$ accepts the norm if:

(1) There is a desire in D which is not fulfilled in $NMAS$, but it is fulfilled in $NMAS'$.
(2) For all other agents $b \in A$, we have that the optimal decision of agent b assuming $\neg\mathbf{V}_{\sim b}$ implies $\neg\mathbf{V}$.

Norms do not always need to be accepted in order to be fulfilled, since the sanction provides a motivation to the agents. However, for a norm to be really effective must be respected due to its acceptance, and not only due to fear of sanctions.

Decision making is not restricted to the normative multiagent system, but it can be extended to the negotiation protocol. We do not consider an equilibrium analysis as in the acceptance relation, but we use the game-theoretic notion called *backward induction*. In the literature on multi-agent systems it is sometimes referred to as *recursive modelling*.

The games we define are as follows. First the agents negotiate the joint goal and the norm, then they make a decision in the normative multiagent system to either fulfill the obligations or accept the associated sanctions.

Definition 8.3. A history h_1 dominates a history h_2 at step i if they have the same set of actions at step $1 \ldots i-1$, are optimal for step $i+1 \ldots$, and $outcomes(h_1) \succ_a outcomes(h_2)$, agent a performing action i prefers the set of possible outcomes of h_1 to the set of possible outcomes of h_2.

A history is optimal at step i if it is not dominated by another history, and it is optimal at all steps $j > i$.

A history is optimal if it is optimal at step 1.

The behavior of agents in the negotiation protocol is illustrated in the following example.

Example 8.1. (Continued) Reconsider Example 6.3 with history h, together with history h' which is like h until $action_3$ while it continues in this way:
$action_4 : propose(a_3, d_3 = \langle \tau''_\delta, \langle s_1, s_2, s_3 \rangle \rangle)$ where
$\tau''_\delta = \{task_1(a_1), task_2(a_2), task_3(a_2), task_4(a_2), task_5(a_3)\}$
$action_5 : accept(a_1, d_3)$
$action_6 : reject(a_3, d_3)$
$action_7 : breakit(a_1)$

Assume that according to agent a_3 both h and h' are optimal from $action_5$ onwards, then agent a_5 compares the two histories to decide how to choose between the two histories. Likewise, the agents of the first three actions decide their optimal actions based on the optimal choice of agent a_3 for $action_4$.

There are a couple of issues with backward induction. First, the length of the histories has to be bounded, because otherwise there is no starting point for the backward induction. Second, if the agents' desires are not common knowledge, then we have to add the desires of agent a_i according to a_j, the desires of agent a_i according to agent a_j according to a_k, *etc.* An efficient representation of such nested modalities is still an open problem. For this reason, in game theory it is usually assumed that the game is common knowledge.

9. Negotiation and related work in normative multiagent systems

In this section we compare norm negotiation in the social delegation cycle with other work on normative multiagent systems.

9.1. *The origin of norms*

Social norms and laws can be used to guide the emergent behavior of multiagent systems.[12] But where do these social norms and laws come from? The following four possibilities have been discussed or suggested in the literature on multiagent systems.

(1) Norms are off-line designed by the agent programmer.[33] However, off-line design has been criticized for open systems.
(2) Norms are created by a single agent in a "legislator" role.[11] The legislator faces the decision problem to decide which norm, sanction and control system has to be created. However, the agent playing the legislator role is vulnerable for attacks or fraud.
(3) Norms emerge spontaneously like conventions.[4,34] However, conventions can be criticized as a coordination mechanism for dynamic systems, because they develop only after a long time (if at all) and they are difficult to change.
(4) Norms are negotiated by (a subset of) the agents in a "democracy", which raises the problem to devise fair and efficient norm negotiation protocols.

Other approaches of norm emergence are based on simulation techniques. Verhagen,[39] for example, considers the problem of the learning of norms. He does not focus however on the acceptance of norms but rather on the influence of normative comments on previous choices. Thus this approach is outside the scope of this paper.

9.2. *Negotiation of obligation distribution*

The social delegation cycle distinguishes norm negotiation from other kinds of negotiation in normative multiagent systems. In contract negotiation agents create new regulative and constitutive norms in a legal context[7] and in negotiation of obligation distribution a group of agents negotiates how to distribute a joint task.[5] These other kinds of negotiation are not based on a social delegation cycle, and therefore give other powers to the agents.

Norm negotiation is related to the problem how a group obligation together with a group sanction can be distributed among the members of the group. In particular, negotiation of obligation distribution[5] is analogous to the second step of the social delegation cycle, extended with a penalty (π) for the agent who breaks the negotiations (which may depend on the agent, e.g., older boys may be punished harder for breaking negotiations than the younger ones). The model distinguishes among three types of sanctions:

- the sanction associated with the group obligation, which is imposed when the obligation is violated, regardless which agent is responsible for it;
- the sanctions associated with the negotiated deal, which are imposed if one of the agents does not fulfill its part of the deal;
- the sanctions associated with the break penalty π.

Moreover, the negotiation is driven by fear of a group sanction, including the sanction for breaking the negotiation. In the model of norm negotiation developed in this paper, the negotiation is driven by the agents' desires. Agents are cooperative in the sense that they have a social goal to achieve.

In some cases it is a drawback to be the only agent able to see to the fulfilment of part of an obligation, but in other cases it may be an advantage, because of the power it gives to the agent over the other agents during the negotiation. This is illustrated by a backward induction argument.

9.3. *Contract negotiation*

As the philosopher of law Ruiter[30] shows, from the legal point of view, legal effects of actions of the members of a legal system are complex and contracts do not concern only the regulative aspects of a legislation (i.e., the rules of behavior specified by obligations), or the constitutive part of it (i.e., the rules introducing institutional facts such bidding in an auction). Rather, contracts are *legal institutions*: "systems of [regulative and constitutive] rules that provide frameworks for social action within larger rule-governed settings".[30] In our model of contract negotiation games,[7] the larger setting is represented by a normative system which establishes the set of possible contracts.

Contract negotiation concerns only two or a few other agents, and therefore there is no need for the three phases of the social delegation cycle, and in particular there is no social goal. Moreover, there is the additional problem not present in social norm negotiation about the choice of contract partners. This choice of contract partners gives a lot of power to agents who have unique or rare abilities. The social delegation cycle however suggests that the powers of agents are less important. In social goal generation, the situation is symmetric for all agents. Only in norm negotiation, like in negotiation of obligation distribution, the powers of agents may be a strategic advantage (or disadvantage!).

9.4. *Normative system as an agent*

In our other work we discuss also applications of normative multiagent systems.[7,8] In those papers the agents consider the normative system as an agent, and they attribute mental attitudes to it, because the agents are playing games with the normative system to determine whether to fulfill or violate norms. We refer to this use of the agent metaphor as "your wish is my command": the goals of the normative agent are the obligations of the normal agents. As an agent with goals, the normative system has to decide whether to consider behavior as a violation and to sanction violations.

In the present paper, however, the agents play games with other agents, and the attribution of mental attitudes to normative system is not a necessary assumption. We thus abstract from the role of the normative system in recognizing violations and sanctioning them: here we assume that violations and sanctions are a direct consequence of the behavior of the agents.

10. Further research

10.1. *Mechanisms*

Within the proposed general model of the social delegation cycle, actual procedures can be defined and studied. In this paper, we propose a fairly general formal model of the social delegation cycle, which delimits the kind of norms that can be created, but that does not give an actual procedure to create norms. The reason is that we aim to capture the fundamental properties of the social delegation cycle, which later can be used to design actual procedures. However, compared to informal characterizations of the construction of social reality, such as in the work of Searle[32], our model is fairly limited as we do not introduce for example beliefs or institutions. Desirable properties may be soundness (compliance with our framework), completeness (for each possible goal there is a goal generated), conciseness of goals and norms generated, generality of goals and norms generated, strictness of goal generation and norm creation, *etc*.

10.2. *Balancing goal negotiation, norm negotiation and acceptance*

We define three steps in the social delegation cycle, i.e., goal negotiation, norm negotiation and acceptance, which do not have to be done sequentially but also can be done iteratively. Moreover, norm negotiation could have been split into obligation negotiation and sanction negotiation. We have not formalized how the elements of the social delegation cycle are balanced. For example, strictly defined norm creation procedures only create norms that will always be accepted, and analogously strictly defined goal generation procedures generate only goals for which a norm can be created that is accepted.

10.3. *Trust*

For more realistic but also more complex social trust, we have to enrich the model with beliefs. We have to extend the merging operators to merging in the context of beliefs.[15] Consequently, we have to introduce beliefs in norm creation, and we have to make the acceptance relation relative to beliefs.

10.4. *The negotiation of permissive norms*

It is not directly clear how the social delegation cycle can explain the creation of permissive norms. One way to proceed is to define permissions as exceptions within hierarchical normative systems.[11] However, also other kinds of permissions have been proposed,[27] and it is unclear which kind of permission should be used in the social delegation cycle.

10.5. *Social institutions and the negotiation of constitutive norms*

How to take social institutions into account in the social delegation cycle? Based on Searle's construction of social reality, we may introduce besides the obligations or regulative norms also constitutive norms, which are definitions of the normative system based on a counts-as conditional.[7]

11. Summary

The social delegation cycle gives cognitive foundations to a social phenomenon, as it relates agent desires to the norms of the society. Individual agents have desires, which turn into social (or joint) goals. A social goal is individualized by a social norm. The individual agents accept the norm, together with its associated sanctions and rewards, because they recognize that it serves to achieve their desires the cycle started with. The cycle thus explains the creation of norms from desires from a rational perspective, because norms are only accepted if they are respected by the other agents, which explains also why sometimes sanctions are needed. In this sense we explain norms in the tradition of Kant, who emphasized the rational aspects of norms.

The social delegation cycle may be seen as a generalization of single agent decision making, in which also the two steps of deriving goals and deriving plans for goals can be distinguished. For example, in the BOID architecture[10] there are components for goal generation and for goal-based planning. Additional issues in the social delegation cycle are the role of sanctions and rewards, the acceptance relation, and the implicit assumption of fairness in goal and norm negotiation. Moreover, though we have not considered this extension in this paper, in the social delegation cycle institutions may play a role. Sanctions may be interpreted either as cues that the other agents see to their task, or as decommitment possibilities for the agents themselves.[31,35] In some cases sanctions must be associated with the norms to ensure that agents fulfill the norm, and therefore to ensure that the agents accept the norm, but in some other cases this is not necessary.

The two following prototypical examples of coordination games and prisoner's dilemma can be represented in our model.[4] First, agents do not want to crash into each other, and the norm to drive on the right side of the road (or the left side, for that matter) is accepted by all members. In this coordination game, no sanction is necessary and the norm may be called a convention. Second, agents want to cooperate in a prisoner's dilemma, so the norm to cooperate is accepted by all members. In this case, a sanction must be associated with the norm, because otherwise the agent will defect (as game theory shows). In this paper we discuss a more complicated example in which there are various kinds of agents, each capable to perform various kinds of task, and they have to negotiate which tasks the group will perform (social goal), and which agents perform which task (the norm). Sanctions are added to avoid free riding.

We formalize the social delegation cycle combining theories developed in a generalization of belief revision called merging operators, planning and game theory. First, we formalize allowed proposals in social goal negotiation as a merging process of the individual agent desires, for which we extend existing merging operators to deal with rules. Second, we formalize allowed proposals in norm negotiation as a planning process for both the obligation and the associated sanctions or rewards. Third, we formalize the acceptance relation as both a belief of agents that the norm leads to achievement of their desires, and the belief that other agents will act according to the norm, introducing a notion of normative equilibrium which states that agents fulfill norms when other agents do so.

References

1. C. Alchourron, P. Gärdenfors, and D. Makinson. On the logic of theory change. *Journal of Symbolic Logic*, 50(2):510–530, 1985.
2. A. Anderson. A reduction of deontic logic to alethic modal logic. *Mind*, 67:100–103, 1958.
3. K.J. Arrow. *Social choice and individual values*. Wiley, New York, second edition, 1963.
4. G. Boella and L. van der Torre. Δ: The social delegation cycle. In *Deontic Logic: 7th International Workshop on Deontic Logic in Computer Science (DEON'04)*, volume 3065 of *LNAI*, pages 29–42, Berlin, 2004. Springer.
5. G. Boella and L. van der Torre. The distribution of obligations by negotiation among autonomous agents. In *Procs. of 16th European Conference on Artificial Intelligence (ECAI'04)*, pages 13–17, Amsterdam, 2004. IOS Press.
6. G. Boella and L. van der Torre. Enforceable social laws. In *Procs. of 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'05)*, pages 682–689, New York (NJ), 2005. ACM Press.
7. G. Boella and L. van der Torre. A game theoretic approach to contracts in multiagent systems. *IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, 36(1):68–79, 2006.
8. G. Boella and L. van der Torre. Security policies for sharing knowledge in virtual communities. *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans*, 36(3):439–450, 2006.
9. G. Boella, L. van der Torre, and H. Verhagen. Introduction to normative multiagent systems. *Computation and Mathematical Organizational Theory, special issue on normative multiagent systems*, to appear.
10. J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.
11. E. Bulygin. Permissive norms and normative systems. In A. Martino and F. Socci Natali, editors, *Automated Analysis of Legal Texts*, pages 211–218. Publishing Company, Amsterdam, 1986.
12. C. Castelfranchi. Modeling social action for AI agents. *Artificial Intelligence*, 103(1-2):157–182, 1998.
13. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213–261, 1990.
14. R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In *Intelligent Agents V (ATAL'98)*, volume 1555 of *LNCS*, pages 99–112. Springer, Berlin, 1998.
15. M. Dastani and L. van der Torre. Specifying the merging of desires into goals in the context of beliefs. In *Procs. of The First Eurasian Conference on Advances in Information and Communication Technology (EurAsia ICT'02)*, volume 2510 of *LNCS*, pages 824–831, Berlin, 2002. Springer.
16. F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7:69–79, 1999.
17. Virginia Dignum, Javier Vázquez-Salceda, and Frank Dignum. A model of almost everything: Norms, structure and ontologies in agent organizations. In *Procs. of 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, pages 1498–1499, New York (NJ), 2004. ACM.
18. M. Esteva, J. Padget, and C. Sierra. Formalizing a language for institutions and norms. In *Intelligent Agents VIII (ATAL'01)*, volume 2333 of *LNCS*, pages 348–366, Berlin, 2001. Springer.
19. M. Esteva, J.A. Rodriguez-Aguilar, C. Sierra, and W.W. Vasconcelos. Verifying norm consistency in electronic institutions. In *Procs. of Workshop on Agent Organizations at AAAI'04*, San Jose (CA), 2004.
20. P. Gärdenfors. *Knowledge in flux*. MIT Press, Cambridge (Massachusetts), 1988.
21. A. Jones and J. Carmo. Deontic logic and contrary-to-duties. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 3, pages 203–279. Kluwer, Dordrecht (NL),

2001.
22. H. Katsuno and A.O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(3):263–294, 1992.
23. S. Konieczny and R.P. Pérez. On the frontier between arbitration and majority. In *Procs. of 8th International Conference on the Principles of Knowledge Representation and Reasoning (KR'02)*, pages 109–120, San Mateo (CA), 2002. Morgan Kaufmann.
24. F. Lopez y Lopez, M. Luck, and M. d'Inverno. Contraining autonomy through norms. In *Procs. of 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'02)*, pages 674–681, New York (NJ), 2002. ACM.
25. D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic*, 29(4):383–408, 2000.
26. D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic*, 30(2):155–185, 2001.
27. D. Makinson and L. van der Torre. Permissions from an input-output perspective. *Journal of Philosophical Logic*, 32(4):391–416, 2003.
28. J. J. Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29(1):109–136, 1988.
29. A. S. Rao and M. Georgeff. The semantics of intention maintenance for rational agents. In *Procs. of 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 704–710, San Mateo, CA, 1995. Morgan Kaufmann.
30. D.W.P. Ruiter. A basic classification of legal institutions. *Ratio Juris*, 10(4):357–371, 1997.
31. T. Sandholm, S. Sikka, and S. Norden. Algorithms for optimizing leveled commitment contracts. In *Procs. of 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, pages 535–541, San Mateo, CA, 1999. Morgan Kaufmann.
32. J.R. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.
33. Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: Off-line design. *Artificial Intelligence*, 73(1-2):231–252, 1995.
34. Y. Shoham and M. Tennenholtz. On the emergence of social conventions: Modeling, analysis and simulations. *Artificial Intelligence*, 94(1–2):139–166, 1997.
35. V. Teague and L. Sonenberg. Investigating commitment flexibility in multiagent contracts. In S. Parsons, P. Gymtrasiewicz, and M. Wooldridge, editors, *Game Theory and Decision Theory in Agent-Based Systems*, pages 267–292. Kluwer, Dordrecht (NL), 2002.
36. M. Tennenholtz. On stable social laws and qualitative equilibria. *Artificial Intelligence*, 102(1):1–20, 1998.
37. L. van der Torre. Contextual deontic logic: Normative agents, violations and independence. *Annals of Mathematics and Artificial Intelligence*, 37(1-2):33–63, 2003.
38. L. van der Torre and Y. Tan. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence*, 27(1-4):49–78, 1999.
39. H. Verhagen. On the learning of norms. In *Procs. of Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'99)*, 1999.
40. G. H. von Wright. Deontic logic. *Mind*, 60:1–15, 1951.