Δ: The Social Delegation Cycle

Guido Boella[1] and Leendert van der Torre[2,3]*

[1] Dipartimento di Informatica - Università di Torino- Italy. E-mail: guido@di.unito.it
[2] CWI Amsterdam - The Netherlands. E-mail: torre@cwi.nl
[3] Delft University of Technology - The Netherlands

Abstract. In this paper we consider the relation between desires and obligations in normative multiagent systems. We introduce a model of their relation based on what we call the social delegation cycle, which explains the creation of norms from agent desires in three steps. First individual agent desires generate group goals, then a group goal is individualized in a social norm, and finally the norm is accepted by the agents when it leads to the fulfilment of the desires the cycle started with. We formalize the social delegation cycle by formalizing goal generation as a merging process of the individual agent desires, we formalize norm creation as a planning process for both the obligation and the associated sanctions or rewards, and we formalize the acceptance relation as both a belief of agents that the fulfilment of the norm leads to achievement of their desires, and the belief that other agents will act according to the norm.

1 Introduction

The relation between obligations and actions is a classical field of study in deontic logic. However, when we consider actions of agents with beliefs and desires, then some questions arise which are traditionally not studied in this area. In agent theory, the relation between desires and obligations has been formalized in BOID agent architectures as a combination of BDI agent architectures [10] and normative (BO) agent architectures, and more generally in normative multiagent systems (NMAS) as a combination of multiagent systems (MAS) and normative systems (NS) for applications like virtual communities [5]. Whereas BDI and NMAS conceptualize the agent's decision making behavior in terms of goals and desires, BO and NS conceptualize the agent's behavior in terms of obligations and permissions.

$$BOID = BDI + BO \qquad NMAS = MAS + NS$$

However, the proposed BOID architectures and normative multiagent systems do not explain how desires and obligations are related. Consequently, the formalization of desires and obligations has raised many questions. For example, is the logic of desire different from the logic of obligation [20]? How do agents deal with conflicts between desires and obligations in their decision making [11]? Why do agents often respect obligations even if they know that their violations are not or cannot be sanctioned? What

* Supported by the ArchiMate research project.

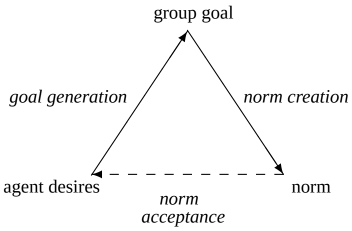

Fig. 1. Δ: the social delegation cycle.

does this imply for the rational creation of norms in such systems, and which mechanisms do not work properly without a normative system? How are social constructions like normative systems constructed from multiagent systems [27]? When is a separation of powers as in trias politica a necessary precondition for norm creation to be efficient?

In this paper we introduce the social delegation cycle, which explains the creation of norms from desires from a rational (e.g., Kantian) perspective. We assume that norms are only accepted if they are respected by the other agents, and therefore sometimes sanctions are needed. Informally, it consists of three steps visualized in Figure 1. Individual agents have desires, which turn into group (or joint, or social) goals. A group 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.

We study the social delegation cycle in a formal framework. The research questions of this paper are:

1. How to balance goal generation, norm creation, and acceptance?
2. How to formalize joint goal generation? We formalize goal generation as a merger of individual desires.
3. How to formalize norm generation? We formalize norm creation as a planning problem, distinguishing between creation of the obligation and creation of the associated sanctions and rewards;
4. How to formalize the acceptance relation? We formalize the acceptance relation by distinguishing between the fulfilment of the agents' desires, and the belief that other agents will fulfill the norm.

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 elsewhere [5, 8, 9], which is based on rule based systems and input/output logics. Moreover, this other work is based on the assumption that the normative system can be modelled as an agent. This paper is not based on this assumption, but it is related to it, as we explain in detail in Section 9.

The layout of this paper is as follows. In Section 2 we discuss the balance between goal generation, norm creation and acceptance. In Section 3 we define the conceptual model in which we study and formalize the social delegation cycle, and in Section 4 we define the logic of rules. In Section 5 we formalize goal generation, in Section 6 we formalize norm creation, and in Section 7 we formalize the acceptance relation.

2 Social delegation cycle

When developing a formal model for the social delegation cycle, we have to make two fundamental choices.

- We may define a general model of the social delegation cycle, defining a range of possibilities, or we may define an actual procedure. The two are not exclusive, since we can first define a general theory of social delegation cycle, thereafter desirable properties within this framework, and finally procedures within the framework that satisfy some or all of the desirable properties.
- We have to define how the elements of the social delegation cycle, i.e., goal generation, norm creation and acceptance, 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.

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 [27], our model is fairly limited as we do not introduce for example beliefs or institutions. This issue is discussed in Section 10.

Concerning the balance between the elements of the cycle, we do not aim to define strict goal generation and norm creation procedures. The reason is that we believe that our setting is more realistic and may cover a wider range of social delegation cycles. Moreover, it facilitates the use of formal theories developed elsewhere, such as merging theories for joint goal generation, planning theories for norm creation, and game theories for acceptance. We consider the definition of strict mechanisms more relevant for the design of mechanisms of norm creation.

Our model builds on several existing formal theories and formalizes the three steps as follows:

Goal generation generates a set of goals based on merging operators, which have been proposed as generalizations of belief revision operators inspired by social choice theory.

Norm creation creates for each goal a set of norms (or revisions of existing norms) based on planning theories as used in most theories in artificial intelligence.

Acceptance relation accepts or rejects a norm based on game theories. We assume that norms are thus only accepted if they are respected, and we formalize the acceptance relation by distinguishing between fulfilment of the agents desire, and the belief that other agents will fulfill the norm.

Before we present our formalizations, we define in the following two sections the conceptual framework we use based on rule based systems, and the logic of rules based on input/output logics.

3 Conceptual model

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

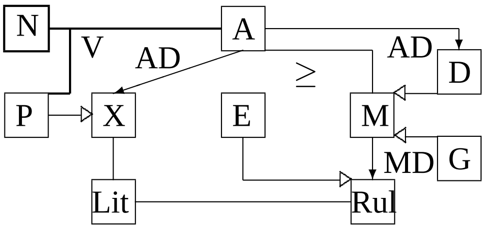

Fig. 2. Conceptual model of normative multiagent system.

The model consists of a set of agents (A), which are described (AD) by a set of boolean variables (X) including *decision variables* it can perform and desires (D) guiding its decision making. The motivational state of the group (G) is composed of its goals. 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 [11]. 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 or goals of other agents 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 1 (AS). *An agent set is a tuple $\langle A, X, D, G, AD, \geq \rangle$, where:*

- *the agents A, variables X, agent desires D and group goals G are four finite disjoint sets. We write $M = D \cup G$ for the motivations defined as the union of the desires and goals.*
- *an agent description $AD : A \to 2^{X \cup D}$ 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 $\geq : A \to 2^M \times 2^M$ is a function from agents to a transitive and reflexive relation on the powerset of the motivations containing at least the subset relation. We write $\geq_a$ for $\geq (a)$.*

Desires and 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, 25]. Agents may share decision variables, or desires, though this complication is not used in this paper. Background knowledge is formalized by a set of effect rules (E).

Definition 2 (MAS). *A multiagent system is a tuple $\langle A, X, D, G, AD, E, MD, \geq \rangle$, where $\langle A, X, D, G, AD, \geq \rangle$ is an agent set, and:*

- *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$(X) = 2^{\mathit{Lit}(X)} \times$ Lit(X), be 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 $MD(S) = \{MD(s) \mid s \in S\}$.*

We now extend the multiagent system to a normative multiagent system to take norm generation into account. To describe the normative system, we introduce a set of norms (N) and a norm description that associates violations with variables (V).

Definition 3 (NMAS). *A normative multiagent system $NMAS$ is a tuple*

$$\langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$$

where $MAS = \langle A, X, D, G, AD, E, MD, \geq \rangle$ is our multiagent system, and moreover:

- *the 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 $V(n, a)$ represents the parameter that counts as a violation by agent a of the norm n.*

We define sanction and reward-based obligations in the normative multiagent system using an extension of Anderson's well-known reduction [2], like Meyer [24] also does: 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. Moreover, we can also have that x is obligatory for agent a while it is a decision variable of another agent b. The logic of obligations, sanctions and rewards satisfies only replacements by logical equivalents.

Definition 4 (Obligation). *Let $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$. 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 $\neg V(n, a) \rightarrow r \in E$.*

The obligation for x is called ought-to-do when $x \in$ Lit$(X \setminus P)$ and it is called ought-to-be when $x \in$ Lit(P).

This is clearly a very weak notion of obligation, and more sophisticated notions within this kind of framework are developed elsewhere [5]. In this paper we now turn to the representation of the desires and goals.

4 Logic of rules

We use a simplified version of the input/output logics introduced in [21, 22]. 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 (Input/output logic [21]). *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*	*(simple-minded reus. 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 and $B \vdash_{iol} p \rightarrow q$ iff $p \rightarrow q \in out(B)$, 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 and output constraints checks whether the derived conditional goals are consistent with the input.

Definition 6 (Constraints [22]). *Let B be a set of rules, and C a set of literals. B is consistent with C, written as $cons(B \mid C)$, iff there do not exist two contradictory literals p and $\neg p$ in $C \cup \{l \mid B \vdash_{iol} C \rightarrow l\}$. We write $cons(B)$ for $cons(B \mid \emptyset)$.*

Due to space limitations we have to be brief on technical details with respect to input/output logics, see [21, 22] for their semantics, further details on their proof theory, the extension with the identity rule, alternative constraints, and examples.

5 Joint goal generation by merging agent desires

We characterize the goal generation process as a merger or fusion of the desires of the agents, which may be seen as a particular kind of social choice process [19]. In this paper, we use the merging operators for merging desires into goals in the context of beliefs, defined in [15]. We adapt these operators in two ways. First we simplify the operators, because we do not use beliefs. Secondly, and most importantly, we make them more complex, because we extend the operators defined on propositional formulas to merge rules.

Definition 7. *A rule base B is a set of rules, a rule set S is a multi-set of rule bases. Two rule sets S_1 and S_2 are equivalent, noted $S_1 \leftrightarrow S_2$, iff there exists a bijection f from $S_1 = \{B_1^1, \dots, B_1^n\}$ to $S_2 = \{B_2^1, \dots, B_2^n\}$ such that $out(f(B)) = out(B)$. We write $\bigwedge S$ for the union of all rules in S, and $\sqcup$ for union with multi-sets.*

Most of these postulates are generalizations of belief revision postulates [1, 16, 18]. (R0) states that the result of merging complies with the integrity constraints. (R1) ensures that, when the integrity constraints are consistent we always manage to extract a coherent piece of information from the knowledge set. (R2) says that, if possible, the result of the merging is simply the conjunction of the knowledge bases of the knowledge set with the integrity constraints, (R3) is the principle of irrelevance of syntax. The purely 'merging' postulates are (R4), (R5) and (R6). (R4) is what is called the fairness postulate. It ensures that when merging two knowledge bases, the operator cannot give full preference to one of them. (R5) and (R6) correspond to Pareto's conditions in social choice theory [3] and were proposed in [26] to model fitting operators. Finally (R7) and (R8) state conditions on the conjunction of integrity constraints and make sure that 'closeness' is well-behaved [18]. 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 8. *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 can be accepted [19], but due to space limitations we do not discuss them. For the same reason we do not discuss the semantics of merging operators. The merging operator is illustrated in the following example.

Example 1. Let $\vdash_{iol}$ be out_3, and consider four rule bases each consisting of a single rule $S = \{\{\top \to p\}, \{\top \to q\}, \{p \to r\}, \{q \to \neg r\}\}$. Now $cons(\bigwedge E)$ does not hold, so due to R1 we cannot have $\nabla_{\emptyset}(S) = \bigwedge S$. They can be merged into a maximal subset of these rules, for example we may have $\nabla_{\emptyset}(S) = \{\top \to p, \top \to q, p \to r\}$ or $\nabla_{\emptyset}(S) = \{\top \to q, p \to r, q \to \neg r\}$. Note that the latter merger selects a maximal consistent subset of S, but it does not select a set of rules that maximizes the output $\{x \mid \nabla_{\emptyset}(S) \vdash_{iol} \top \to x\}$ (a distinction discussed in [22]). If we assume $\vdash_{iol}$ be out_1, then $cons(\bigwedge E)$, and due to R1 we have $\nabla_{\emptyset}(S) = \bigwedge S$.

In our conceptual model, goals are a subset of the merger of desires of the agents.

Definition 9 (Goal generation). *There is a rule merging operator* ∇ *such that* $MD(G) \subseteq \nabla_E(MD(D_x) \mid x \in A)$.

In the latter definition we use variable x to refer to agents. We use variables also in many other places, e.g., in the following example, but these variables are just used to shorten the presentation and are not part of the logical language. It is just some *syntactic sugar*. For example, quantification over rules means that it is schema: there is a set of rules, one for each agent involved. Since the set of agents A is finite, we are still in propositional logic. Joint goal generation is illustrated by the following example.

Example 2. Let $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V\rangle$ with the following ingredients:

variables in X:
$\{\neg collision(x_1, x_2), accident, drive\_right(x), drive\_left(x) \mid x_1, x_2, x \in A\}$.
Moreover, each agent can decide to drive on left or right side of street, e.g., $X_a = \{drive\_right(a), drive\_left(a)\}$,

effect rules E:
$\{drive\_right(x_1) \wedge drive\_left(x_2) \to collision(x_1, x_2) \mid x_1, x_2 \in A\}$
$\cup \{(\bigwedge_{x_1,x_2 \in A} \neg collision(x_1, x_2)) \to \neg accident\}$
$\cup \{collision(x_1, x_2) \to accident \mid x_1, x_2 \in A\ \}$
$\cup \{collision(x_1, x_2) \to collision(x_2, x_1) \mid x_1, x_2 \in A\}$
If two agents do not drive on same side then they collide, and if there are no collisions then there is no accident

desires: $D_x = \{\top \to \neg collision(x, y) \mid y \in A\ \}$ for each agent $x \in A$. Agents desire not to be part of a collision.

goal $G = \{\top \to \neg accident\}$.

The system generates a joint goal of $NMAS$ for absence of accidents.

Goals can be generated using negotiation processes. Alternatively, the process can be facilitated by an agent playing the role of legislator. Here we do not further consider the construction of goals.

6 Norm creation

We formalize norm creation as a planning problem, distinguishing between the creation of the obligation and the creation of the associated sanctions and rewards. In some cases sanctions must be associated with the norms to ensure that some agent fulfills the norm, and therefore to ensure that the other agents accept the norm, but in some other cases this is not necessary. Here are two prototypical examples.

- 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 case, no sanction is necessary and the norm may be called a convention. Other examples of this kind can be found in coordination games in game theory.
- 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).

The two elements of norm creation are formalized as two sequential steps: first determining the obligation, and thereafter determining the associated sanctions or rewards. The first step is essentially a planning problem: the obligations of the agents must imply the joint goal $Y \rightarrow g$. 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, we introduce in context Y an obligation $Y \wedge \sim\!x \rightarrow V(n, a)$. Roughly, the condition is that all obligations x imply the goal g.

However, to determine whether the obligations imply the 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 goal g will be achieved, we propose the following condition: if every agent fulfills its newly introduced obligation, and it fulfills all its other obligations, then g is achieved. We define a global violation constant **V** as the disjunction of all indexed violation constants like $V(n, a)$, i.e., $\mathbf{V} = \bigvee_{n \in N, a \in A} V(n, a)$.

Definition 10 (Norm creation). *Let $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V\rangle$ with $Y \rightarrow g \in MD(G)$. The parameters contain the global violation constant $\mathbf{V} \in P$ and E contains the following set of rules:*
$\{V(n, a) \rightarrow \mathbf{V} \mid n \in N, a \in A\} \cup \{\neg\mathbf{V} \rightarrow \neg V(n, a) \mid n \in N, a \in A\}$
The creation of norm n' to achieve joint goal $Y \rightarrow g$ leads to the updated normative multiagent system $\langle A, X, D, G, AD, E \cup E', MD, \geq, N \cup \{n'\}, V\rangle$ such that:

1. *The norm n' is not already part of N;*
2. *A set of rules $E' = \{Y \wedge x \rightarrow V(n', a) \mid a \in A, x \in Lit(X)\}$ 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 joint goal is satisfied;*
3. *$cons(E \mid Y \wedge \neg\mathbf{V})$, it is possible that no norm is violated.*

The creation of norms is illustrated by the following example.

Example 3. Let $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$ as defined in Example 2. Assume that the normative system creates a norm n' with the following obligations: $\forall a \in A : \neg right\_side(a) \rightarrow V(n', a)$: $\neg right\_side(a)$ counts as a violation of norm n' by agent a.

The second step is adding sanctions and rewards. The condition of this second step is that sanctions are disliked, and rewards are desired.

Definition 11 (Norm creation with sanctions and rewards). *Let $NMAS =$ be a normative multiagent system $\langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$ with $Y \rightarrow g \in MD(G)$. The creation of norm n' with sanctions and rewards to achieve joint goal $Y \rightarrow g$ leads to the updated system $\langle A, X, D, G, AD, E \cup E' \cup E'', MD, \geq, N \cup \{n'\}, V \rangle$ with:*

1. *The creation of norm n' to achieve joint goal $Y \rightarrow g$ leads to updated system $\langle A, X, D, G, AD, E \cup E', MD, \geq, N \cup \{n'\}, V \rangle$ and*
2. *The set of rules $E'' = \{Y \wedge V(n', a) \rightarrow s \mid a \in A, s \in Lit(X)\} \cup \{Y \wedge \neg V(n', a) \rightarrow r \mid a \in A, r \in Lit(X)\}$ 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.*

Sanctions are illustrated by the following example.

Example 4. Let $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$ with the following ingredients:

agents A**:** $\{a, b\}$;
variables in X**:**
$\{c(a), c(b), cooperation, s(a), s(b)\}$ with $X_a = \{c(a)\}$, $X_b = \{c(b)\}$, each agent can cooperate (e.g., $c(a)$) or not, each agent can be sanctioned (e.g., $s(a)$) or not.
effect rules E**:**
$\{c(a) \wedge c(b) \rightarrow cooperation\}$, there is cooperation if both agents cooperate.
desires D**:** $D_a = \{\top \rightarrow \neg c(a), \top \rightarrow cooperation, \top \rightarrow \neg s(a)\}$,
$D_b = \{\top \rightarrow \neg c(b), \top \rightarrow cooperation, \top \rightarrow \neg s(b)\}$.
Agents desire to defect (e.g., $\neg c(a)$), but they also desire cooperation, and they desire not to be sanctioned.
goal $G = \{\top \rightarrow cooperation\}$, the system has generated a joint goal of $NMAS$ for cooperation.

Assume that the normative system creates a norm n' with the following obligations: $E' = \{\neg c(a) \rightarrow V(n', a) \mid a \in A\}$: $\neg c(a)$ counts as a violation of norm n' by agent a. Moreover, it adds the following sanctions: $E'' = \{V(n', a) \rightarrow s(a) \mid a \in A\}$.

There may be a third step that adds controls to the obligations, sanctions and rewards. We do not consider this extension in this paper.

7 Norm acceptance

An agent accepts a norm when the obligation implies the desires 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 fulfillment leads to fulfillment of its personal desire the cycle started with. This implies that fulfillment of the goal g is kind of normative equilibrium.

Definition 12 (Decision). *Let* $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$. *The optimal decision of agent* $b \in A$ *given a set of literals* C *is defined as follows.*

- *The set of decisions is the set of subsets of* $Lit(X_b)$ *that do not contain a variable and its negation. A decision* δ *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* $b \in A$ *are the desires whose body is part of the decision, but whose head is not.*
  $U(\delta, b) = \{d \in D_b \mid MD(d) = L \to l, E \vdash_{iol} C \cup \delta \to l'$ *for* $l' \in L$ *and* $E \not\vdash_{iol}$ $C \cup \delta \to l\}$.
- *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 $\mathbf{V}_{\sim b}$ of the global violation constant $\mathbf{V}$ 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 13 (Acceptance). *Let* $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$, *and let* $NMAS' = \langle A, X, D, G, AD, E \cup E' \cup E'', MD, \geq, N \cup \{n'\}, V \rangle$ *be the system after the creation of a norm and its associated sanctions and rewards. The parameters contain the global violation constants* $\mathbf{V}_{\sim b} \in P$ *and* E *contains the following rules:*
$\{V(n, x) \to \mathbf{V}_{\sim b} \mid n \in N, x \in A \setminus \{b\}\} \cup \{\neg \mathbf{V}_{\sim b} \to \neg V(n, a) \mid n \in N, x \in A \setminus \{b\}\}$
An agent $a \in A$ *accepts the norm if:*

1. *There is a desire in* D *which is not satisfied in* $NMAS$*, but it is satisfied 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.

It can easily be shown that in the two running examples, both norms are accepted.

8 Further research

8.1 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, see [15]. Consequently, we have to introduce beliefs in norm creation, and we have to make the acceptance relation relative to beliefs.

8.2 The creation 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 [12].

8.3 Social institutions and the creation 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 [9].

9 Related work

9.1 Other work

The relation between 'desires' or internal motivations and 'obligations' or external motivations has been studied in many areas, for example:

Religion. The Golden Rule or the ethic of reciprocity is found in the scriptures of nearly every religion. It is often regarded as the most concise and general principle of ethics. It is a condensation in one principle of all longer lists of ordinances such as the Decalogue.

Ethics. Kant's categorical imperative [17] expresses the moral law as ultimately enacted by reason and demanding obedience from mere respect for reason.

Political theory. Marx (ideology) [23]: the ruling class forms a theory (obligations) justifying itself (its desires).

Social theory. Norms (obligations) are only accepted if the legislator does not make them only for his own interests (desires) ([14]).

Agent theory. Your wish is my command: the desires of the master are the obligations of the slave.

Within formal and semi-formal agent theory, there has been some work by Castelfranchi, Conte and colleagues on norm adoption and norm acceptance [14].

9.2 Normative system as an agent

In other work we discuss applications of normative multiagent systems [5], of which the formal machinery based on rule based systems and input/output logics has been developed in various papers. 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. 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. In our other work, we have informally discussed the notion of the social delegation cycle in a short paper [4]. In that short paper we have suggested that the social delegation cycle can be used to explain the agent metaphor "your wish is my command", because the group goal from which the norm is created, may be interpreted as the goal of the normative system, and the normative system is doing a kind of planning.

In this framework, we have not discussed the merging of desires into group goals, but we have mentioned the notion of rational norm creation in a second short paper [7]. In that paper we do not present a formalization of norm creation, and we do not consider norm creation within the context of the social delegation cycle. Finally, we introduce an extension of our formal model with constitutive norms in [9] and we observe that constitutive norms play an important role in norm creation, but we do not formally study it. The creation of permissions in this framework has been mentioned in [6].

Finally, in none of our other work we have discussed the acceptance relation, and we have not discussed games between ordinary agents.

10 Summary

In this paper we consider the relation between desires and obligations in normative multiagent systems. We introduce a model of their relation based on what we call the social delegation cycle, which explains the creation of norms from agent desires in three steps. First individual agent desires generate group goals, then a group goal is individualized in a social norm, and finally the norm is accepted by the agents when it leads to the fulfillment of their initial desires. The social delegation cycle may be seen as a generalization of single agent decision making, which can also be defined as a combination of goal generation and 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 generation. Moreover, in the social delegation cycle institutions may play a role.

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 joint goal generation as a merging process of the individual agent desires, for which we extend existing merging operators to deal with rules. Second, we formalize norm creation 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.

There are two main directions for further research. First, the theories have to be extended with beliefs and institutions to cover social delegation cycles based on trust and norm creation by institutions. Second, for the social delegation cycle efficient mechanisms should be designed which can be employed in actual implementations of normative multiagent systems. 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, *et cetera*.

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. Attributing mental attitudes to normative systems. In *Procs. of AAMAS'03*, pages 942–943. ACM Press, 2003.
5. G. Boella and L. van der Torre. Local policies for the control of virtual communities. In *Procs. of IEEE/WIC Web Intelligence Conference*, pages 161–167. IEEE Press, 2003.
6. G. Boella and L. van der Torre. Permissions and obligations in hierarchical normative systems. In *Procs. of ICAIL'03*, pages 109–118, Edinburgh, 2003. ACM Press.
7. G. Boella and L. van der Torre. Rational norm creation: Attributing mental attitudes to normative systems, part 2. In *Procs. of ICAIL'03*, pages 81–82, Edinburgh, 2003. ACM Press.
8. G. Boella and L. van der Torre. Contracts as legal institutions in organizations of autonomous agents. In *Procs. of AAMAS'04*, New York, 2004.
9. G. Boella and L. van der Torre. Regulative and constitutive norms in normative multiagent systems. In *Procs. of 9th International Conference on the Principles of Knowledge Representation and Reasoning*, Whistler (CA), 2004.
10. M.E. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, Harvard (Massachusetts), 1987.
11. 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.
12. 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.
13. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
14. R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In *Intelligent Agents V (ATAL'98)*, volume 1555 of *LNAI*, pages 319–333. Springer, 1999.
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. Springer, 2002.
16. P. Gärdenfors. *Knowledge in flux*. MIT Press, Cambridge (Massachusetts), 1988.

17. I. Kant. *Kritik der praktischen Vernunft*. Johann Friedrich Hartnoch, Riga, 1788.
18. H. Katsuno and A.O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.
19. 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, Toulouse, 2002.
20. J. Lang. Conditional desires and utilities - an alternative approach to qualitative decision theory. In *Procs. of ECAI'96*, pages 318–322, Budapest, 1996.
21. D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.
22. D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic*, 30(2):155–185, 2001.
23. K. Marx. *Das Kapital*. Verlag von Otto Meissner, Hamburg, 1867.
24. 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.
25. 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, Montreal, 1995.
26. P.Z. Revesz. On the semantics of arbitration. *International Journal of Algebra and Computation*, 7(2):133–160, 1997.
27. J.R. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.