This paper was published in *Social Order in Multiagent Systems*, edited by R. Conte and C. Dellarocas, Kluwer Academic Publishers, Boston etc., 2001.

CHAPTER 7: SOCIAL INSTITUTIONS, NORMS, AND PRACTICES

Wolfgang Balzer, University of Munich LMU

and

Raimo Tuomela, University of Helsinki

Abstract:

We submit a model of social institutions which binds together the two central components of institutions, a) a 'behavioral' system of social practices as repeated patterns of collective intentional actions and b) the normative *Überbau* consisting of a task-right system which on the one hand is influenced and in basic cases even induced by the 'underlying' practices and on the other hand serves to stabilize them. An explicit and relatively simple connection in terms of sanctions is drawn between actions which are obligatory or permitted by special positions on the one hand and the 'ordinary' course of actions which occurs in social practices within an institution on the other hand. Obligations and rights are not simply bound to actions, but to systems of actions given in the form of systems of social practices. This adds an essential component which has been neglected in formal treatments so far. The inclusion of social practices yields a rich structure in which the emergence and maintenance of norms can be tackled in a realistic way.

INTRODUCTION

The need for clarifying the interplay of actions and norms within social institutions is keenly felt among social scientists and in the multi-agent community. In sociology, the mainstream approach to institutions is in game theoretic terms, e.g. (Schotter, 1981), but there also are approaches

using a power structure (Balzer, 1990), (Coleman, 1974), or stressing the cognitive level (Conte & Castelfranchi, 1995). In game theory the representation of actions and expectations is very idealized and far away from application to comprehensive real-life institutions. In the power centered approach so far the intentional, normative part has remained at an informal level. In AI, the study of cooperation has included organizational features (Durfee et al., 1987), (Prietula et al., 1998) and norms (obligations, 'social laws') (Barbuceanu, 1997), (Moses & Tennenholtz, 1995), and has led to formal accounts of institutionalized power, norms, rights, and obligations (Jones & Sergot,1997). One main restriction of these accounts is their lack of reference to the mental sphere of attitudes. This prevents the exploitation of attitudes as a means of governing action.

A comprehensive theory of institutions is still missing which makes explicit the overall macro structure, the norms, and the systems of actions as well as the interplay between these components. These features have to be formalized so that a comprehensive model may guide further fine grained studies which can lead to implementations. We submit a model of social institutions in the sense of institutional organizations. This model captures both the normative and the action component. It binds together a) a 'behavioral' system of social practices as repeated patterns of collective intentional actions and b) the normative *Überbau* consisting of a task-right system, which on the one hand is influenced and in basic cases even induced by the 'underlying' practices and on the other hand serves to stabilize them. The model is not fully general in that we leave corporate actors and some aspects of jointness out of consideration.

The present model makes precise two special features which are missing in previous attempts. First, an explicit connection in terms of sanctions is drawn between actions which are obligatory or permitted by special positions on the one hand and the 'ordinary' course of actions which occurs in social practices within an institution on the other hand. Though this connection has been discussed for quite some time (e.g. Pörn, 1970), it has not received the manageable formalization needed for computer applications. The new feature of our model is that obligations and rights are not simply bound to actions, but to systems of actions given in the form of systems of social practices. This adds an essential component which has been neglected so far (but see (Balzer, 1990)). The inclusion of social practices yields a rich structure in which

the emergence and maintenance of norms can be tackled in a realistic way. Second, an institution requires particular attitudes with a complex content refering to the whole institution. Roughly, these are mutual beliefs with the content 'all members behave according to the institution's norms'. These contents are spelled out in detail, using the structure which is given to an institution and its included norms.

The model thus offers a fresh start by making explicit the interplay between actions, the attitudes which are at work in triggering them, and the system of rights and obligations which stabilizes the system of social practices (actions) in an institution. We believe that the model yields a realistic basis for detailed case studies,[1] and also for subsequent studies of the emergence of task-right systems.

1. STATES AND ACTIONS

Our model is a state space model in which the states are sets of sentences, indexed by a time variable. The states are relativized to individuals or groups, so that we can describe different states in which different persons or groups find themselves at the same time. States need not be closed under implication and no consistency requirements are made.

To make the state change approach philosophically and theoretically justified, some qualifications are needed - see (Tuomela & Sandu, 1994) and (Tuomela, 1995) for discussion. Here we will directly proceed to our logical model. Actions are modelled as changes of state. Any pair (C, E) of sets of sentences of a given language L describes a potential transition from a 'previous' state C to a subsequent, state E. The sentences occurring in C and E must be such that under the right conditions they could describe some real action. In this case C describes a state in which the conditions for the action are satisfied, and E describes a state in which the effect of the action obtains. We distinguish between a) action types (C, E) for which the elements of C and E are formulas of L possibly containing variables, b) potential actions for which members of C and E must be sentences (closed formulas) and c) actions which really are performed. The latter are represented by $\mathit{perf}(t, i, (C, E))$, reading 'at time t, individual or group i performs the action described by (C, E)'. An action (C, E) at t may fail to produce its effect E (see below).

[1] However, even a simple example is beyond the space available here.

For a set C of formulas we write $C[t,i]$ and $C[t,i_1,...,i_n]$ to denote the set of sentences obtained from C by replacing all variables by the names t,i, resp. $t,i_1,...,i_n$ for instants and persons. To economize on notation we also write $C[t,i]$ if i denotes a group, $i=\{i_1,...,i_n\}$. For the actual performance of an action we assume that the names occurring in the sets $C[t,i]$, $E[t,i]$ are the same that are used in the first two arguments of the *perf* predicate: $perf(t,i,(C[t,i],E[t,i]))$, and we simply write $perf(t,i,(C,E))$ and, still more simply, $perf(t,i,C,E)$. Also, we will abbreviate actions $(C[t,i],E[t,i])$ by $a[t,i]$, and we agree that whenever a occurs in a sentence containing some expression $perf(t,i,a)$ then a is an abbreviation for $a[t,i]$.

A primitive A is used in order to pick out those pairs (C,E) representing action types from the set of all pairs (C,E) of sets of formulas. From A, a set A^* of (descriptions of) potential actions can be defined in terms of closure. A^* contains all pairs (C^*,E^*) such that, for some $(C,E)\in A$, C^*,E^* are the sets of closures of formulas in C and E. If $S(L)$ and $F(L)$ denote the sets of sentences and formulas of a language L, we thus distinguish between a) transition types[2] $(C,E)\in \mathbf{po}(F(L))\times\mathbf{po}(F(L))$, b) action types $(C,E)\in A$, c) potential actions $(C,E)\in A^*$, and d) actions $perf(t,i,C[t,i],E[t,i])$.

The sentences in $S(L)$ will also be used in order to express the content of some mutual belief held among the persons considered which is central and constitutive for a social institution. Roughly, this content expresses that all members in the institution behave according to the tasks and rights assigned to them by their respective positions in the institution.[3] As this content comprises a major part of the structure of an institution, the sentences in $S(L)$ must be rich enough to express this structure.

2. FRAMES

The conceptual arena in which we will talk about actions, rights, obligations, social practices and institutions we call a *frame*. A frame is built

[2] $\mathbf{po}(X)$ denotes the power set of set X.

[3] This is of course an idealized picture, which still is central for understanding the normative content of an institution, thus what would happen in a normatively ideal world. In actual life of course violations occur and norms are followed unintentionally or for the wrong reasons.

up from

- a non-empty, finite set J of individuals or persons
- a finite, non-empty set G of groups such that $G \subseteq \mathbf{po}(J)$ and each $g \in G$ has at least two elements (we use I as an abbreviation for $J \cup G$)
- a non-empty, finite set ATT of attitude kinds containing at least belief, intention, and goal
- a finite, linear order $(T, <)$, representing time
- a finite set O of 'ordinary objects'
- a language L with sets $S(L)$ and $F(L)$ of sentences and formulas
- a set $A \subseteq \mathbf{po}(F(L)) \times \mathbf{po}(F(L))$ of descriptions of action types
- a function x: $T \times I \rightarrow \mathbf{po}(S(L))$, the state function
- a function *caus*: $T \times \mathbf{po}(S(L)) \times T \rightarrow \mathbf{po}(S(L))$, the causal function
- a relation *perf* $\subseteq T \times I \times A^*$, the relation of actual performance
- a relation *catt* $\subseteq T \times G \times ATT \times A^*$ expressing collective attitudes (e.g. collective goals and intentions, mutual beliefs)
- a relation *incom* $\subseteq A^* \times A^*$ of incompatibility of potential actions
- a relation $ex \subseteq T \times J$ ('existence')
- a relation $sanc \subseteq \{+, -\} \times A \times A$ ('sanctions').

A frame basically consists of a state space, the states of which are described by sets of sentences (members of $S(L)$). The development of states over time is represented by the state function x which is relativized to individuals or groups. The sentences in $x(t, i)$ describe the state in which individual or group i is at time t. For each non-maximal instant t, the 'next' instant is denoted by $t + 1$. $caus(t, X, t')$ denotes the effect at time t' caused by the presence of X at t. The 'cause' here is described by the sentences in X. If these sentences are satisfied at t, then the cause X is present at t. At t' the ensuing effect is $caus(t, X, t')$, $caus(t, X, t') \subseteq \cup_{i \in I} x(t', i)$.

perf $(t, i, C[t, i], E[t, i])$ reads: at t, i performs (or the members of i collectively perform) action $(C[t, i], E[t, i])$. For proper individuals $i \in J$ this comprises individual action and for groups $i \in G$ collective action. An action may fail in the sense that for all subsequent t', $E[t, i] \not\subseteq caus(t, C[t, i], t')$.

catt represents collective attitudes in the distributed sense. $catt(t, g, att, a)$ means that, at time t, the members of group g share the we-

attitude of kind att with content a.[4] Briefly, actor i in group $g = \{i_1, ..., i_n\}$ has the we-attitude with content a, $we\text{-}att_i(a)$, iff $att_i(a) \wedge bel_i(\forall j \in g(att_j(a)) \wedge mubel(\forall j \in g(att_j(a))))$, where $att_i(a)$ means that i has the individual attitude of kind att that a, and $mubel$ means mutual belief. Sharing a we-attitude in a group means that all members in the group share it, or in a weaker sense that a fixed percentage of them share it. For example, the distributed collective intention of actors $\{i, j\}$ to perform some action a means that the two intend to do their respective parts of a, believe that the respective Other intends to do his part, believe that the respective Other belives that 'I' intend to do 'my' part, and so on, see e.g. (Balzer & Tuomela, 1997), (Tuomela, 2000), (Wooldridge & Jennings, 1997) for accounts of distributed collective attitudes.

$incom(a, b)$ expresses that the potential actions a and b are incompatible. This may be much weaker than inconsistency, incompatibility may simply be due to practical reasons. Note that $incom$ cannot operate at the level of action types because in many cases incompatibility only arises when two actions are performed at the same time.

$ex(t, j)$ means that at t, individual j exists as an active member. For each $g \in G$ and each t, we denote by g_t the set of members of G existing at t, $g_t = \{i \in J / ex(t, i)\}$. We assume that for each $j \in J$ there exist t^l_j, t^u_j such that $t^l_j < t^u_j$ and for all t with $t^l_j \leq t < t^u_j$, $ex(t, j)$, and in t^l_j and t^u_j are the 'smallest' and 'largest' such instants, i. e. j has an uninterrupted period of existence.

$sanc$ is used to express that an action type b is a sanction for another action type a. As every sentence is a formula, $sanc$ also can be applied to actions so that we can speak of action b being a sanction for action a. We distinguish between sanctions of the form $(+, a, b)$ representing a sanction b following the performance of a, and sanctions of the form $(-, a, b)$ in which b is a sanction for a not having been performed.[5] We say that i's action $a[t, i]$ at t is *sanctioned* iff there is another agent j performing an action $b[t', j]$ at some later time t' such that b is a sanction

[4] In ordinary language one would say 'the group has that attitude'. However, the precise meaning of this phrase is still under discussion so that we here work with the technically established notion of a shared we-attitude of some relevant kind, see (Tuomela, 2000).

[5] As $sanc$ in the following will be applied only to actions we need not bother about the precise interpretation of 'action type a not having been performed'.

of a: $\exists b \in A^* \exists j \exists t'(t < t' \wedge (+, a, b) \in sanc \wedge perf(t, i, a) \wedge perf(t', j, b))$. Similarly i's not doing a at t is sanctioned iff not $perf(t, i, a)$ and there are b, j and $t' > t$ such that $(-, a, b) \in sanc$ and $perf(t', j, b)$. Sanctions here are always understood in the negative sense.

In A we may distinguish between action types (and potential actions) involving one or more individuals. Action types (C, E) satisfying[6] $\forall t, i$: $perf(t, i, C, E) \rightarrow \exists j \in J(i = \{j\})$ are called *individual*, those which do not satisfy this condition being called *collective* action types.[7] By CA and IA we denote the sets of collective and individual action types.

A *frame* y thus has the form $y = (J, T, ATT, O, G, <, L, A, x, caus, perf, catt, incom, ex, sanc)$.

3. SOCIAL PRACTICES

A social institution consists of two central parts, an 'underlying' system of social practices and a (weakly) normative *Überbau*. We analyzed single social practices in (Balzer & Tuomela, 2003).

A social practice roughly is a repeated pattern of collective action in which a collective attitude[8] of kind *att* (usually belief or intention) with content B is formed in a group, and an action of a corresponding action type (C, E) is then performed. In general, the relation between content B and action type (C, E) may be opaque, but in the present first analysis we assume that both are identical, i. e. $B = (C, E)$. For example, if the attitude kind is *intention*, the group may repeatedly form the collective intention 'we have sauna together next Saturday' and perform the collective action of having sauna together each 'next' Saturday. Both the content 'we have sauna together next Saturday' and the corresponding action are represented in the format (C, E) of an action type where C contains sentences like 'the sauna is operative', 'most persons in the group are healthy' etc., and E contains sentences like 'sufficiently many

[6] The formula says that all agents i_r participating in the action ($i_r \in \{i_1, ..., i_n\}$) are identical with j.

[7] This does not guarantee of course that actions of such collective action types are 'collective' in any interesting sense of this term.

[8] Compare Sec. 2.

persons meet at 10 a.m. in the lobby', 'the persons enter the sauna and bath' etc.[9]

Slightly modifying the account in (Balzer & Thomela, 2003), the core of a social practise is given by three items:

- a kind att of attitude
- a content (C, E) of that attitude such that
- (C, E) is a collective action type.

By a collective action type we only mean a type which is realized by a 'collective' of several persons, in contrast to individual action types, the actions of which can be performed by one person. In a frame $y = (J, T, ATT, O, G, <, L, A, x, caus, perf, catt, incom, ex, sanc)$ we assume that $att \in ATT$ and $(C, E) \in A$.

To these core items we add functions describing trigger conditions for attitudes ($trigatt$) and actions ($trigact$) which are specific for the particular action type (C, E) under consideration and are represented by sets of formulas. If all the trigger conditions in these sets are instantiated and true this will lead to the formation of the collective attitude, and to the subsequent performance of a collective action of type (C, E). In the sauna case, a trigger condition for the attitude might be, for example, that the persons call each other to see whether they will have company, and a trigger condition for action will be that it is Saturday, 10 a.m. Our notion of trigger condition is a deterministic one. For indeterministic context it needs to be relaxed, e.g. probabilistically - see the discussion in (Balzer & Thomela, 2003).

Moreover, we use numerical functions suc for the $success$ of a collective action, and thr to specify a threshold. The value of suc is increased or decreased depending on the success of the performance of the action, and the constant thr gives a threshold. If the success function drops below the threshold for several successive repetitions of the practice, the practice is likely to terminate.

Each formation of the collective attitude followed by a corresponding action and the latter's causal effects take place in one period $z = (t_1, ..., t_4)$ in which four points of time are distinguished. At the first point t_1 the trigger conditions for the attitude are present, at t_2 the

[9] See (Balzer & Tuomela, 2003) for a detailed analysis and more elaborate examples.

collective attitude is formed, at t_3 the corresponding action is executed, and at t_4 the causal effects of that action are noted. In a social practice such a four step pattern is repeated over and over, so we consider a sequence of periods $(z^i)_{i=1,2,3,....}$. By P^* we denote the set of all periods z^i pertaining to a given social practice.

In a frame y a *social practice with core* $(g, att, (C, E))$ now can be defined as a system $(g, att, (C, E), (z^i)_{i=1,2,3,...}, trigatt, trigact, suc, thr)$, where $g \in G$ is a group, att a kind of attitude, (C, E) a collective action type, $(z^i)_{i=1,2,3,...}$ a sequence of periods and[10]

- $trigatt : T \times \{g\} \times \{att\} \times A \to \mathbf{po}(S(L))$,
- $trigact : T \times \{g\} \times \{att\} \times A \to \mathbf{po}(S(L))$,
- $thr : \{g\} \times \{att\} \times A \to \mathbf{N}$,
- $suc : P^* \times \{g\} \times \{att\} \times A \to \mathbf{N}$.

Moreover, some axioms have to assure that the four step schema described above is repeated over a sufficiently large number of periods. In particular, we assume the following.[11]

A1) The sequence (z^i) of periods is embedded into the overall time structure $(T, <)$ such that the periods 'follow' each other. At the different points of time t the active members of group g are those found in g_t.

A2) In each period and at each specified instant t of that period, g_t contains 'sufficiently many' members so that the characteristic action type (C, E) can be performed.

A3) If the collective attitude with content $a[t, g_t]$ is present in the group at t (among the active members g_t) then the trigger conditions for action will lead at the next instant $t + 1$ to the action's $a[t, g_t]$ being performed 'because of' that attitude, and conversely, if at $t + 1$, $a[t + 1, g_{t+1}]$ is performed because of the attitude then, at t, the trigger conditions must have been present.[12]

A4) If in the first instant of a period the trigger conditions for the attitude with content (C, E) obtain for the active members of group g and the success level for actions of the kind (C, E) is above the threshold, then the collective attitude will be formed and be present at

[10] N is the set of natural numbers.

[11] See (Balzer & Tuomela, 2003) for formal details.

[12] $t + 1$ need not be chosen according to the pattern of instants in the periods. We assume that attitudes persist in the sense of (Cohen & Levesque, 1990).

the second point of time in that period, and conversely, if the collective attitude is present at the second instant, the trigger conditions for that attitude must be present in the first instant and the success level in the first instant must be above the threshold.

A5) At the end of each period the success function is updated as follows. If the action performed in that period was a success, the function value is increased by one, otherwise it is decreased by one.

'Success' is expressed by reference to the action description (C, E). The action is successful if its effects, E at $t+1$, in fact, are among the causal consequences of its conditions C at t ($E \subseteq caus(t, C, t+1)$).[13] In figure 1) the causal flow characteristic for a practice with core $(g, att, (C, E))$ is shown for one period $z = (z_1, z_2, z_3, z_4)$ (possible variations of membership in g being suppressed).

Fig.1

$$
\begin{array}{ll}
z_1 & x(g, z_1) \wedge trigatt((C, E), z_1) \subseteq x(g, z_1) \wedge thr \leq suc(z) \\
 & \quad \Big| \; caus\,, \text{ A2} \\
 & \quad \downarrow \\
z_2 & catt(g, att, (C, E), z_2) \wedge trigact(z_2) \subseteq x(g, z_2) \\
 & \quad \Big| \; caus\,, \text{ A1} \\
 & \quad \downarrow \\
z_3 & perf(z_3, g, C, E) \\
 & \quad \Big| \; caus \\
 & \quad \downarrow \\
z_4 & caus(z_3, C, z_4) \;\; (E \subseteq caus(z_3, C, z_4)?)
\end{array}
$$

At the first instant z_1 the set of individuals g is in state $x(g, z_1)$. If in that state the trigger conditions for the attitude att with content (C, E) are satisfied and the attitude had been sufficiently successful, this will causally lead to the presence of the attitude $catt(g, att, (C, E), z_2)$ at the next instant z_2 (this is part of the content of axiom A2). If at z_2 the trigger conditions for action corresponding to att and (C, E) are satisfied this will causally lead to the performance of such an action $perf(z_3, g, C, E)$ in the next period z_3 (this is part of the content of A1).

[13] Using slightly different formulations of these axioms, in (Balzer & Tuomela, 2003) necessary and sufficient conditions are stated for the 'survival' of a practice over time.

A third causal transition then produces the result $caus(z_3, C, z_4)$ of that action which may be different from the effect E specified by (C, E).

In order to define a *system* of several different social practices we use a set SP of *names* for social practices, and a function f which to each (name of a) social practice assigns a value $(g, att, (C, E))$ specifying the group g, the kind of attitude att and the action type (C, E) specific for that practice (its core).

D1 s is a *system of social practices* iff $s = (J, T, ATT, O, SP, <, L, A, x, caus, perf, catt, incom, ex, sanc, f)$ and
1) $y = (J, T, ATT, O, G, <, L, A, x, caus, perf, catt, incom, ex, sanc)$ is a frame
2) SP is a finite, non-empty set (of labels of social practices)
3) f: $SP \rightarrow G \times ATT \times CA$ and $\cup\{\pi_1(f(sp))/sp \in SP\} = J$
4) for all $sp \in SP$ and all g, att, a, if $f(sp) = (g, att, a)$ then in y there exists a social practice with core (g, att, a).

We do not require that different practices in a system of practices be compatible though this assumption makes good sense in most institutions, and in particular in organizations whose task-right system is officially specified.

4. OBLIGATIONS AND RIGHTS

In an institution, obligations and rights are attached to the positions *pos* which the persons occupy in it. Each person *holds* a specific position *pos* which we identify with two sets of action types, $pos = (OB_{pos}, RI_{pos})$, $OB_{pos} = \{o_1, ..., o_m\}$, $RI_{pos} = \{r_1, ..., r_n\}$ such that holders of *pos* are obliged to perform actions of types $o_1, ..., o_m$ and have the right to perform actions of types $r_1, ..., r_n$. Obligations and rights thus are represented in the following way. Person i in position *pos* is *obliged* to do a iff a is one of the action types occurring in OB_{pos} and the conditions for executing a obtain. Briefly, an obligation to do a is represented by '$a \in OB_{pos}$' for some position *pos* in the institution. Similarly, a right to do a in position *pos* is represented by '$a \in RI_{pos}$'.

Using the format (C, E) for action types, with conditions C and effects E, and the state function x and performance relation *perf* described earlier, this representation of rights and obligations can be connected

with actions in a natural way. Consider some person i in position pos, and some action type $o = (C, E)$ obligatory for pos, i.e. $o \in OB_{pos}$. If i is in a state $x(t, i)$ in which the conditions for o are satisfied ($C[t, i] \subseteq x(t, i)$) then i should perform $o[t, i]$. At the non-normative level 'i should perform $o[t, i]$' corresponds to 'if i does not perform $o[t, i]$ then i gets sanctioned': $\neg perf\,(t, i, o) \rightarrow \exists j \exists t' \exists b(t < t' \wedge\ perf\,(t', j, b) \wedge sanc(-, o, b))$. In the case of rights the connection is a bit more complicated. If $r = (C, E)$ is covered by a right of i ($r \in RI_{pos}$ and $holds(t, i, pos)$) and i is in a state in which she could perform r ($C[t, i] \subseteq x(t, i)$) then no other person j should perform any action b interfering with r. That is, for any other person j and action $b[t, j] = (C'[t, j], E'[t, j])$ which j could perform at time t ($C'[t, j] \subseteq x(t, j)$), and which is incompatible with r ($incom(r[t, i], b[t, j])$), j should not perform $b[t, j]$ at t. Again, 'j should not perform $b[t, j]$ at t' at the non-normative level corresponds to 'if j performs $b[t, j]$ at t then j's action $b[t, j]$ at t gets sanctioned': $perf\,(t, j, b) \rightarrow \exists k \exists t' \exists c(t < t' \wedge\ perf\,(t', k, c) \wedge sanc(+, b, c))$.

This account provides a relatively simple connection between the normative level, the normative force of obligations and rights, and the level of actions and sanctions. It thus might serve as a basis for further investigations of how and why obligations and rights emerge and are upheld.

The action types attached to rights and obligations are anchored in a system of social practices as follows. We assume that each such action type comes from one of the practices in an 'underlying' system of practices, i.e. the action type is 'part of' the core of such a practice. This assures that no contrived actions figure in the rights and obligations. Rights and obligations are concerned only with socially entrenched action types. We cannot assume, however, that an action type expressing, say, an obligation, is simply identical with the action type of a social practice, for the latter describes a collective action while the former describes an individual one. In order to bridge this gap we use a relation *part* between collective actions (or action types) and their individual *parts* writing $part((C, E), i, (C_i, E_i))$ to express that (C_i, E_i) is an individual action (type) which forms person i's part of the collective action (type) (C, E). A part (C_i, E_i) need not be unique; a person i may have several parts to perform in the collective action (C, E).[14]

[14] Of course, this covers up all the problems of spelling out the individual parts

D2 tr is a *task-right system for* the system s of social practices $(J, T, ATT, O, SP, G, <, L, A, x, caus, perf, catt, incom, ex, sanc, f)$ iff there exist $POS, part$ and $holds$ such that $tr = (POS, part, holds)$ and

1) for all pos, $pos \in POS$ iff there exist $o_1, ..., o_n, r_1, ..., r_m$ such that $pos = (OB_{pos}, RI_{pos})$, where $OB_{pos} = \{o_1, ..., o_n\} \subseteq IA$ and $RI_{pos} = \{r_1 ..., r_m\} \subseteq IA$
2) $part \subseteq CA \times J \times IA$
3) $holds \subseteq T \times J \times POS$
4) for all pos, t, i, if $holds(t, i, pos)$ then $ex(t, i)$
5) for all $pos = (OB_{pos}, RI_{pos}) \in POS$, all $(C, E) \in OB_{pos} \cup RI_{pos}$, all $i \in J$ and all $t \in T$, if $holds(t, i, pos)$ then there exist (C^*, E^*) and $sp \in SP$ such that

5.1) $f(sp) = (g, att, (C^*, E^*))$
5.2) $part((C^*, E^*), i, (C, E))$.

The action types $(C, E) \in OB_{pos}$ are those which holders of position pos are *obliged* to perform (under the right conditions). Whenever the conditions C are satisfied for a person i holding position pos (i.e. $C[t, i] \subseteq x(t, i)$) then i is obliged to perform an action of type (C, E). Action types a in RI_{pos} specify the rights of persons holding position pos. D2-4 says that if agent i holds position pos at time t then i must exist (be an active member of the institutions) at t. Condition D2-5 is depicted in figure 2.

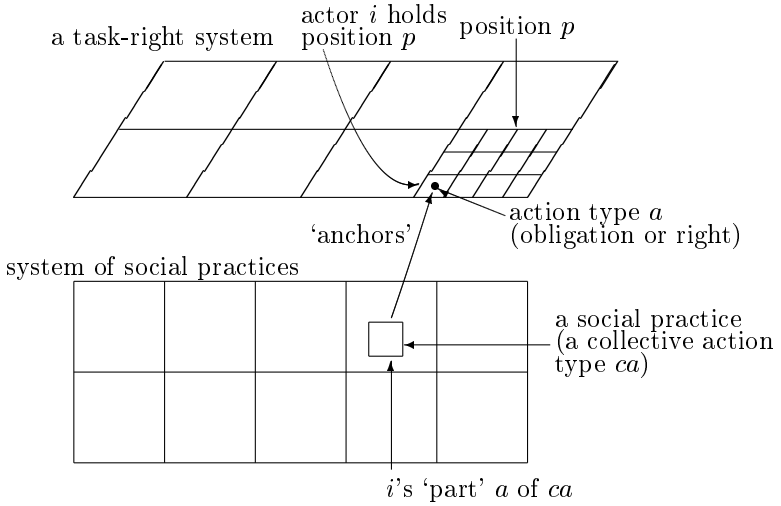
In the normative system on the top each large box represents a position which in turn consists of action types (the small boxes), one of which, $a = (C, E)$, is picked out. In the system of social practices at the bottom each large box depicts a social practice of which only the pertaining collective action type ca is depicted. Decomposing ca into its individual parts, i.e. those individual action types that have to be performed in order to produce a realization of a collective action of that type, we obtain a set of individual action types at the bottom, one of which is depicted by the small box. The condition of entrenchment in 5) now says that each individual action type a on the top is identical with ('comes from', 'is constituted by') one of the individual actions types at

---

of a collective action, and of constructing collective actions out of individual ones. However, for practical purposes it can be assumed that a collective action in fact is constituted by individual, 'basic' actions in the way of dynamic logic, i.e. by recursively forming bigger actions of the form $a \parallel b$ and $a; b$ out of simpler ones, see (Harel, 1984), (Sandu & Tuomela, 1996).

the bottom. Note that this yields a very strong, core notion of an institution in which all norms must relate to 'living' practices. In reality there are many 'parasitic' institutions which draw (part of) their normative system from other institutions.

Fig.2



In such cases there may be normative action types which do not come from any of the underlying practices. Note further that our formulation leaves room for the development of new practices which are not normatively covered. By contrast, an extension of the normative system must by preceded by corresponding extensions at the level of practices.

Using a weak negation of action ('it is not the case that i performs a'), inflating the number of obligations, and assuming some kind of consistency of the task right system we can express the usual connection between rights and obligations as follows. If $a \in RI_{pos}$ and $holds(t, i, pos)$ then for all a^* of type a, all b and all j: if $incom(a^*, b)$ and $holds(t, j, pos')$ then among the obligations of pos' there is one obliging j not to perform b ('if i has the right to do a then every j has the obligation to refrain from actions incompatible with a'). Conversely, if $a \in OB_{pos}$ then there is no right (in the system) of performing an action incompatible with a.

5. SOCIAL INSTITUTIONS

A social institution now consists of a system of social practices plus a task-right system for it. The system of tasks and rights on the one hand normatively mirrors certain combinations of collective action as found in the system of social practices. On the other hand, the normative task-right system by its obligations and rights provides external reasons of institutional action. We submit three axioms. The first, D3-3, is a central, analytic condition. It states that among the members of an institution there is a common belief (*mubel*)[15] that everybody behaves according to the obligations and rights attached to his position. The other two hypotheses are of a contingent, empirical nature, and aim at explaining the role of the normative system. D3-4 and 5 say that people 'usually' perform the actions they are obliged to perform, and 'usually' refrain from actions conflicting with the rights of other members. 'Usually' has to be understood in a statistical way, refering to the numbers of performances and the weights of the different actions and types.[16]

In order to formulate these regularities, let us define, for $a = (C, E) \in A$, and $pos \in POS$, the numbers

- $exopp(a, pos)$, the number of *execution opportunities* of a in pos, as the number of $(t, i) \in T \times J$ such that $holds(t, i, pos) \wedge C[t, i] \subseteq x(t, i)$
- $exec(a, pos)$, the number of *executions* of a in pos as the number of number of $(t, i) \in T \times J$ such that $holds(t, i, pos) \wedge C[t, i] \subseteq x(t, i) \wedge perf(t, i, (C, E))$
- $freq(a, pos)$, the *frequency* of executions of a in pos, by $exec(a, pos)$ / $exopp(a, pos)$
- $vio(a/pos)$, the number of actions *conflicting* with a in pos as the number of $(t, i, j, b) \in T \times J \times J \times A$ such that $holds(t, i, pos)$ and $incom(a[t, i], b[t, j])$ and $perf(t, i, a[t, i])$ and $perf(t, j, b[t, j])$.

Note that in *exopp*, $C[t, i] \subseteq x(t, i)$ need not lead to action, the trigger

[15] See (Balzer & Tuomela, 1997), (Colombetti, 1993) or (Wooldridge & Jennings, 1997) for accounts of mutual belief.

[16] In order to avoid the mutual beliefs in D3-3 to be irrational, given the probabilistic formulations of D3-4 and 5, we should rather use an approximate version of D3-3, too. However, as this would involve substantial additional formalism, we prefer to stick to the simpler, somewhat problematic formulation.

conditions also must occur.

D3 x is a *social institution in force* iff there exist soc and tr such that $x = (soc, tr)$ and

1) soc is a system of social practices
2) tr is a task-right system for y
3) for all $t \in T$: $mubel(t, J, p)$ where $p = p_1 \wedge p_2$ is the following sentence
   $p_1 \equiv \forall j \in J \forall pos \in POS \forall t \in T \forall (C, E)$
   if $pos \in POS \wedge (C, E) \in OB_{pos} \wedge C[t, j] \subseteq x(t, j) \wedge holds(t, j, pos)$
   then $perf(t, j, C, E)$, and
   $p_2 \equiv \forall i, j \in J \forall pos \in POS \forall (C, E) \in RI_{pos} \forall t \in T \forall (C^*, E^*) \in A$, if
   $holds(t, j, pos) \wedge C[t, j] \subseteq x(t, j) \wedge C^*[t, i] \subseteq x(t, i) \wedge perf(t, i, (C^*, E^*))$
   $\wedge\, incom((C[t, j], E[t, j]), (C^*[t, i], E^*[t, i]))$ then i gets sanctioned
4) for all $pos = (OB_{pos}, RI_{pos}) \in POS$ and all $a \in OB_{pos}$,
   $freq(a, pos)$ is close to 1
5) for all $pos = (OB_{pos}, RI_{pos}) \in POS$ and all $a \in OB_{pos}$,
   $vio(a/pos)$ is close to 0.

Sentence p expresses that all members behave (in the social practices) according to their positions (tasks and rights). p_1 says that whenever the conditions of an action type to which i is obliged in her position obtain then i will perform an action of that type. p_2 expresses that all persons can act according to their rights. If another person i performs some action incompatible with j's potential action $(C[t, j], E[t, j])$ to which j is entitled $((C, E) \in RI_{pos} \wedge holds(t, j, pos))$ then i gets sanctioned. These are of course the ideal versions of proxy formulations. Figure 3 shows the overall picture that must obtain if these requirements are satisfied (if the distinction between action types and actions is suppressed).

The circle depicts the array of normatively admitted actions, and the inner rectangle the array of actions really performed. The intersection of these sets represents actions which accord with the given norms, while the difference of the two sets contains those actions which are violations of norms. By D3-4 the relative size of the intersection should be close to 1 while by D3-5 the size of the difference set should be small. Axiom D3-3 requires the inclusion depicted by the arrow at the level of mutual belief. People believe that the realized actions conform to the norms.

A social institution cannot exist or be in force without 'we-mode' thinking and acting, viz. thinking and acting appropriately as a group

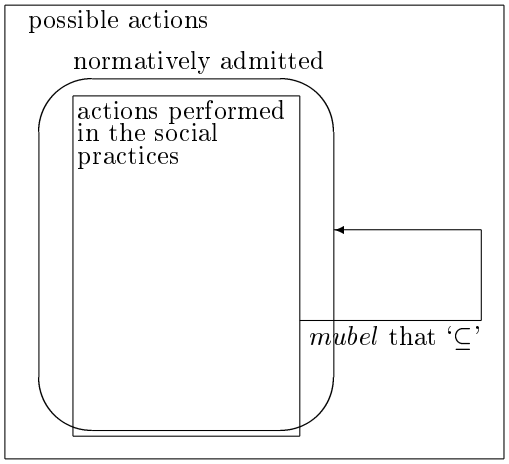
member functioning in the institution.[17] This is important for understanding what an institution is - namely something existing for the group for the use of the group and something which is man-made (intentionally or - in some cases - non-intentionally). An institution involves some basic goals for the institution and it involves a task-right system concerned with the achievement of those goals. Thus there is acting in the right way and acting in the wrong way in an institution. Acting in the right way in its fullest sense involves not only performing the normatively specified actions but also performing them for the right reason, viz. for the reason that they are appropriate in view of the task-right system of the institution. This requirement shows up e.g. in the mutual belief concerning sentence p in D3. This belief must be a we-mode belief specifying that p is for the 'use' of the group members or participants in the institution. (This reason need be only a presupposition reason, and this does not require that the agents in normal circumstances reflect on it.) Unless there is a substantial amount of such action for the right reason there is not the right kind of understanding of the institution in question and, furthermore, the institution will not function well. As to the functionality point, if the functioning of the institution is externally disturbed, then appropriate changes in the social practices and perhaps also in the task-right system may be required. Such changes, however, cannot rationally be made without understanding the nature of the institution and thus the notion of acting in a position for the right reason. The rational design, redesign and change of social institutions also in the case of artificial collectives (such as 'robot societies') thus also must rely on the idea of we-mode thinking and acting in the group, where the we-modeness takes into account the presupposition that an institution is available for the group in question and that the group is at least to some extent committed to its institution.

Hypotheses D3-4 and 5 above are formulated as parts of the definition of the notion of a social institution. Any system qualifying as an institution must satisfy these requirements. The justification for this is that a system in which D3-4 and 5 are not satisfied, a system in which nearly all obligations and rights are violated, cannot be called a social institution. One may want to separate the empirical aspects covered by D3-4 and 5 from a purely conceptual definition which is free from

[17] See (Tuomela, 2000), Chapters 2 and 6 for the notion.

empirical contingencies. On an alternative account which draws a sharp distinction between empirical and conceptual matters, D3-4 and 5 would be removed from the definition and would be read as external criteria for the extent to which an institution (defined by D3-1 to 3) is 'in force' or well functioning.

Fig.3



Finally, we want to point out a difficulty that arises when we reformulate the model keeping syntax and semantics separate in the usual way. In such a setting the sentence p in D3 expressing the mutual belief would contain variables ranging over sets of sentences, like C, E, C^*, E^*, and over pairs of sets of pairs of sentences, like *pos*. Defining validity in such a setting would be a formidable task. The present, set-theoretic approach avoids this at the cost of loosing the explicit, deductive part. At least in the beginning however, this loss seems to be bearable in view of the cost of having syntax separated.

We are indebted to Rosaria Conte for clarifying comments.

REFERENCES

Balzer, W. (1990). A Basic Model of Social Institutions. *Journal of Mathematical Sociology*, 17, 1 - 29.

Balzer, W. and Tuomela, R. (1997). A Fixed Point Approach to Collec-

tive Attitudes. In (Holmström-Hintikka and Tuomela, 1997), pp. 115 - 42.

Balzer, W. and Tuomela, R. (2003). Intentions and the Maintenance of Social Practices. *Autonomous Agents and Multi-Agent Systems* 6, 7 - 33.

Barbuceanu, M. (1997). Coordinating Agents by Role Based Social Constraints and Conversation Plans. Proceedings AAAI-97, 16 - 21.

Cohen, P. R. and Levesque, H. J. (1990). Intention is Choice with Commitment, *Artificial Intelligence* 42, 213 - 263.

Coleman, J. S. (1974). *Power and the Structure of Society.* New York: Norton.

Colombetti, M. (1993). Formal Semantics for Mutual Belief. *Artificial Intelligence*, 62, 341 - 53.

Conte, R. and Castelfranchi, C. (1995). *Cognitive and Social Action.* London: VCL.

Durfee, E. H., Lesser, V. R., Corkill, D. D. (1987). Coherent Cooperation Among Communicating Problem Solvers. IEEE *Transactions on Computers*, 36, 1275 - 91.

Harel, D. (1984). Dynamic Logic. In D. Gabbay and F. Günthner (eds.), *Handbook of Philosophical Logic, Vol.* II , Dordrecht: Reidel, pp. 497 - 604.

Holmström-Hintikka, G. and Tuomela, R. (eds.). (1997). *Contemporary Action Theory. Vol.* 2, Dordrecht: Kluwer.

Jones, I. A. J and Sergot, M. (1997). A Formal Characterization of Institutionalized Power. In E. G. Valdez et al. (eds.). *Normative Systems in Legal and Moral Theory*, Berlin: Duncker and Humblodt, pp. 349 - 67.

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

Pörn, I. (1970). *The Logic of Power.* Oxford: Blackwell.

Prietula, M., Carley, K., Gasser, L. (eds.). (1988). *Simulating Organizations: Computational Models of Institutions and Groups.* Cambridge MA: MIT Press.

Sandu, G. and Tuomela, R. (1996). Joint Action and Group Action Made

Precise. *Synthese*, 105, 319 - 345.

Schotter, A. (1981). *The Economic Theory of Social Institutions.* Cambridge: UP.

Tuomela, R. (1995). *The Importance of Us.* Stanford, Stanford University Press.

Tuomela, R. (2000). *Cooperation: A Philosophical Study.* Philosophical Studies Series, Dordrecht: Kluwer.

Tuomela, R. and Sandu, G. (1994). Action as Seeing to it that Something is the Case. In P. Humphries (ed.), *Patrick Suppes: Scientific Philosopher*, Vol. 3, Dordrecht: Kluwer, pp. 193 - 221.

Wooldridge, M. and Jennings, N. R. (1997). Formalizing the Cooperative Problem Solving Process. In (Holmström-Hintikka and Tuomela, 1997), 143 - 161.