Comput Math Organiz Theor (2006) 12:181–204
DOI 10.1007/s10588-006-9543-9

# Self-regulation through social institutions: A framework for the design of open agent-based electronic marketplaces

Christian Hahn · Bettina Fley · Michael Florian

© Springer Science + Business Media, LLC 2006

**Abstract** In this paper, we argue that allowing self-interested agents to activate social institutions during runtime can improve the robustness (i.e., stability, reliability, or scalability) of open multiagent systems (MAS). Referring to sociological theory, we consider institutions to be rules that need to be activated and adopted by the agent population during runtime and propose a framework for self-regulation of MAS for the domain of electronic marketplaces. The framework consists of three different institutional types that are defined by the mechanisms and instances that generate, change or safeguard them. We suggest that allowing autonomous agents both the reasoning about their compliance with a rule and the selection of an adequate institutional types helps to balance the trade-off between the autonomy of self-interested agents and the maintenance of social order (cf. Castelfranchi, 2000) in MAS, and to ensure almost the same qualities as in closed environments. A preliminary report of the evaluation of the prototype by empirical simulations is given.

**Keywords** Institutions · Self-regulation · BDI · Electronic market · Multiagent systems

C. Hahn (✉)
Department of Deduction and Multiagent Systems, German Research Center for Artificial Intelligence (DFKI), Stuhlsatzenhausweg 3 (Building 43), 66123 Saarbrücken, Germany
e-mail: Christian.Hahn@dfki.de

B. Fley · M. Florian
Department of Technology Assessment, Hamburg University of Technology, Schwarzenbergstr. 95, 21073 Hamburg, Germany
e-mail: bettina.fley@tu-harburg.de

M. Florian
e-mail: florian@tu-harburg.de

## 1. Introduction

The design and development of robust and efficient open multiagent systems (MAS), where a vast amount of heterogeneous agents with different goals, different rationales and varying perceptions of appropriate behavior can interact, is an area of increasing importance in MAS research, especially in the context of Internet applications like electronic marketplaces.

In accordance with Wooldridge, Jennings and Kinny (1999), robustness is the ability of a system to maintain safety-responsibilities even in case of disturbances. Relating to Schillo et al. (2001), robustness criteria regarding open multiagent-based electronic marketplaces are attributes like scalability, flexibility, resistance and agent drop-out safety that can be measured by the relationship between certain safety-responsibilities (i.e., domain oriented *performance criteria*) and domain specific *perturbation scenarios*.

Factors affecting those qualities can be divided into two groups. Firstly, the model of the electronic marketplace itself can cause technical problems reducing the system's performance since the technical realization of open MAS depends on agent interoperability, communication protocols, and reliable infrastructures. Secondly, environmental influences (e.g., demand for new products, newly participating provider agents) and interaction dynamics in the market can cause *interaction outcomes* as well as *system states* that are both undesirable from the perspective of user purposes (reliable and efficient electronic trade) and diminishing the trustworthiness and acceptance of such MAS applications.

Nevertheless, the provision of a clear definition at design-time which actions lead to undesirable outcomes, either on interaction or system level, and the development of appropriate mechanisms to prevent these behaviors are rather difficult, especially if agents are considered to be autonomous regarding their belief, desires, intentions and plans, being able to change their behavior continuously. With respect to markets, three different types of behaviors leading to unfavorable outcomes can be distinguished:

- Agents in markets are generally not benevolent, but self-interested, assumed to maximize their utility by changing goods. According to economic theory, the pursuit of self-interest in markets is not harmful per se with respect to individual and overall outcomes. In contrast, markets are assumed as mechanism that combines the pursuit of individual interests with the achievement of collective interests (highest overall outcome). However, classical economics only took the pursuit of interest by negotiating prices into account. Deception and fraud (in contrast to the honest pursuit of self-interest) may further the utility of single agents, but damage others and is considered to lead to "sub-optimal" market outcomes. Hence, *deception and fraud* can be clearly defined as unwanted behavior. Nevertheless, in large distributed environments like the Internet, the breach of e-contracts or other forms of unreliable trading cannot be prevented completely.
- Other forms of self-interested behavior, which are actually essential in markets, can also cause undesirable outcomes on interaction and system level, but only under certain conditions. For example, the adaptation of prices can also be used to increase market shares or to gain competitive advantages in periods of low demand, leading to *ruinous competition*, insolvency and market break down instead of realizing efficient

market coordination. Moreover, the ability to form organizations may be abused for cartelization or monopolization by powerful entities that try to achieve enough power to set prices and attain higher producer rents. While the formation of organizations improves robustness (Schillo, Knabe and Fischer, 2004b), a certain degree of market concentration may reduce the efficiency and flexibility of the whole marketplace. Nevertheless, the formation of organizations and price adaptations cannot be defined as deviant in general.

- In human markets, agents try to improve their competitive positions depending on the structure of competition, their own positions and the demand situation. Due to a lack of knowledge about the state of the market, agents may choose unfeasible or inadequate strategies with respect to the state of the market and other agents' choices of strategies, leading to a poor performance of the agents themselves, but to unacceptable system states. This also applies to agents-based markets. Local knowledge and information asymmetries may lead to unequal goals and varying perceptions of options (plans) to fulfill self-imposed goals of autonomous agent, impairing the coherence of market interaction as well.

In Distributed Artificial Intelligence (DAI), a common approach to handle these kinds of problems is to resort to a *trusted third party* (cf. Froomkin, 1997; Boella and Damiano, 2002) that establish conventions and norms that standardize interactions, establish safeguards and guarantee that certain intended actions actually take place and unwanted situations are prevented. In contrast, sociology emphasizes phenomena like customs, norms and laws, summed up by the term *institution*, with respect to the question of how actions of human agents are structured and regulated. As a common denominator, differing sociological theories define institutions as rules that are binding, because they provide meaning to agents, guiding their expectations and beliefs, and have a certain obligatory effect claiming validity and prevalence (cf. Esser, 2000). From this perspective, we criticize that the cited work in DAI focuses to much on third parties that safeguard rules rather than on the question how rules can be generated during runtime, being institutionalized as belief forming, prevalent, and obligatory. Admittedly, sociological metaphor of "social institutions" already has been adopted by DAI. However, it has been primarily employed in the sense of social agreements upon interaction styles and shared meanings. State-of-the art research on electronic institutions largely deals with the institutionalization of transaction protocols (cf. Colombetti, Fornara and Verdicchio, 2002; Dignum, 2001, 2004; Esteva et al., 2001).[1]

Therefore, this paper aims at exploiting this sociological metaphor more comprehensively to provide a theory of flexible institutions allowing agents in an electronic marketplace to activate rules (institutions) and to activate or form instances (third parties that generate and safeguard rules) during runtime in order to dynamically self-regulate non-desirable interactions and system states, using different mechanisms to gradually restrict the agent's autonomy with respect to its of internal states (i.e., the agent's control over its behavior and decision making process with respect to the its

[1] Some work (cf. Axtell, 2001) also uses the term institution synonymously with organizations. Note that the sociological use of the term institutions rules that are not capable of acting (cf. Scott, 2001; Esser, 2000), whereas organizations are defined as social entities that consist of members, (material) resources and rules and are able to act as corporative actor.

beliefs, desires and intentions) and external constraints (available options for actions, imminent consequences of actions due to reactions of instances other agents). Using the term autonomy, we refer to *norm autonomy* proposed by Verhagen (2000) specified by the ability to autonomously generating goals based on a system of norms. In Section 3, we show how institutions (the normative component of the agent's model) may influence the agent's deliberation on the action, plan and goal level.

## 2. A sociological perspective on institutions

In sociology, institutions are defined as rules. In contrast to the usage of the term *rule* in computer science, social institutions are not considered to be 'built-in constraints' of human societies that clearly define the actions that are allowed or forbidden and the consequences in case of rule violation. Although explicit instructions are an important property of certain types of social institutions, institutions provide stability of social life not only by constraining, but by guiding agents regarding their options of action, the selection of the appropriate option, and the respective consequences. Besides clear instructions, social rules or institutions consist of social agreements about meanings, taken-for-granted assumptions and appropriate frameworks of action that are often ambiguous and fragmented incoherent in human societies (cf. DiMaggio, 1997).

With respect to the robustness of MAS, the question raises whereon the potential of institutions rests to generate shared assumptions, structure expectations and regulate the actions of agents, particularly because rules have no capacity to act like individuals (agents), collective and corporative agents (organizations, instances). And secondly, this involves the question how this potential can be used for the design of MAS-based open electronic marketplaces. Giving an answer to the first question requires a more detailed definition of the term rule from a sociological perspective, a more detailed explanation what causes the binding nature and a certain claim of validity and prevalence of these rules.

### 2.1. Definition of institutions as rules

In sociology, rule exists that can be summarized by two different conceptions of the word. On the one hand, a rule can be an underlying *principle of action*, defining which interactions are desired, which are unwanted or even forbidden in certain contexts and under certain conditions. Rules in this sense are available to the awareness of agents and can be more or less consciously mastered by them in terms of reflecting about rule violation or rule conform behavior of themselves or others. On the other hand, specific collective behaviors can show a *rule-like character*, i.e., a certain regularity, but do not presuppose a rule as a guiding principle of action. Rules in the sense of regular behavioral patterns are considered to be produced by the aggregate of individual actions of agents that have been confronted with similar structural constraints or environmental states (cf. Bourdieu, 1990: 60f.) and therefore developed similar orientations towards certain objectives of action (goals or desires) as well as shared meanings and assumptions (plans) about how certain things can be done under specific circumstances (cf. Berger and Luckmann, 1966). Social institutions are rules in both senses.

Consequently, not any observable or cognitively available pattern of behavior is an institution. Institutions are considered to be social macro-phenomena that are *durable* and having a *scope in the social space* that reaches beyond informal relationships and temporarily limited encounters among dyads or triads of individuals.

This spatial and temporal scope confers institutions a certain transintentional character meaning that these regular patterns exist largely independent of the will of single agents. Institutions are not at the disposal of single agents, so that they appear as an external constraint. However, the externality of institutions is to a large extent caused by other factors: some authors (cf. Berger and Luckmann, 1966) argue that this partial independence of institutions from agents' intentions due to mere routine and customization. Institutions provide agents with solutions to specific problems of action, i.e., shared meanings, knowledge, and patterns of how things can be done. The more those meanings and patterns become taken-for-granted certainties and are disseminated within a population, the more alternatives will be ignored or considered to be not feasible. However, other important factors for the externality of institutions are possibilities to impose sanctions on rule violating agents, either by collective moral disrespect or physical force. Especially rules that are available to the consciousness of agents facilitate discourses about which behaviors conform to and which violate certain institutions.

Moreover, despite the externality of institutions, agents' actions do not necessarily correspond to collective behavioral patterns. Agents may violate rules. To explain why agents commit themselves to act according to rules, it is not sufficient to refer to external factors. In order to oblige agents, rules necessitate to be accepted as legitimate by the agents themselves. However, *commitment* can have varying origins. Firstly, agents may adopt a rule as a taken for granted certainty (cf. Berger and Luckmann, 1966). Secondly, agents may attribute a rule a certain value of its own (e.g., because a rule is meant to safe-guard collective goods or public welfare). Finally, the commitment may be due to agents own interests, depending whether rule conformity or violating behavior is useful to reach one's goals or to avoid disadvantages (e.g., bad reputation, legal sanctions).

## 2.2. Three types of institutions

This definition already suggests that diverse types of institutions can be distinguished, depending on the degree to which they are available to the awareness of agents, their socio-spatial scope, their degree of durability, externality and capability to commit agents, and in general the strength with which they claim prevalence and validity. Moreover, the previous section already anticipated that institutions are varying with respect to their capability to further, prevent or stop certain actions and hence, differ to the degree to which they restrict the autonomy of agents. In order to develop a framework for MAS that leaves as much autonomy (with respect to the agent's controls over its behavior) as possible to agents, it is necessary to provide a typology of different institutions that identifies which type of institution is required for the regulation of the different, in the introduction specified perturbation scenarios. Following Scott (2001), who discriminates three elements of institutions, we distinguish between three types of institutions with regard to the degree to which they restrict the agents' autonomy and to which they provide solutions for the regulation of the perturbation scenarios mentioned. However, in the remainder of this paper, we mainly refer to the work of sociologist Pierre Bourdieu on rules, regular and regulated behavior (cf. Bourdieu, 1990) for two reasons: firstly, Bourdieu's habitus concept provides insights (cf. Schillo et al., 2000) that help to develop an agent model that enables self-interested agents to reason about their obedience to rules (cf. Section 3). Secondly, the field concept (cf. Schillo et al., 2004a, b) provides a concept that allows analyzing driving forces of institutionalization processes, i.e. sources of institutional practice beyond individual actors (cf. Section 2.3). However, space restrictions do not allow a more detailed summary of those two concepts in this paper.

### *2.2.1. Practical institutions (PI)*

With this type, we refer to observable patterns of collective behavior that are not produced by the consciously managed obedience to a consciously available rule (cf. Bourdieu, 1990: 60). Instead, those patterns (or so-called strategies) result from the actions of agents (1) which try to accumulate different sorts of capital that are accredited in a certain social context (field), e.g., reputation, economic profit, (2) which are confronted with the same structural constraints (similar competitive positions), and (3) which share similar dispositions of perception, reasoning and action (a similar habitus). Both, the generation and the durability of those institutions result to a large extent from the similarity of the agents' dispositions (habitus), which in turn have been acquired by agents through their long-term experiences in a certain social context (social field) and are conditioned by the similarity of competitive positions in those fields. Hence, the socio-spatial scope of those regularities is mainly restricted to a certain agent class, i.e., to those agents sharing similar competitive position over a longer period and hence a similar habitus. As a consequence, the collective behavior of a specific class manifests itself in a certain collective style of action that is recognizable by other agents. Although stylization allows the recognition of a certain behavioral pattern as class specific, practical institutions are no formulated rules. Therefore, the agents' commitment is not influenced by possible sanctions. Rather, commitment towards a style is due to its feasibility with respect to the agents' goals and to the identification of an agent with a certain style, leading to conclusive actions regarding that style. The advantage of this type of institution consists in providing different classes of agents with behavioral patterns of feasible actions and hence, in contributing to the coherence of interaction between agents by learning practical strategies in the electronic marketplace.

### *2.2.2. Normative institutions (NI)*

With this type, we refer to rules that Bourdieu calls quasi-juridical principles (cf. Bourdieu, 1990: 60) and which also can be defined as norms that indicate behaviors that are acknowledged as honorable and morally approved. Norms are more or less formulated and consciously manageable. The socio-spatial scope of those norms is not restricted to specific agent classes, but to communities of agents who share certain values. Norms allow shared judgments of certain actions either as honorable or dishonorable/immoral. In contrast to classes, communities are not defined by the similarity of the agents' social positions, but by network-like relationships between agents that are characterized by trust and commitment towards each other. Therefore, the durability and validity of those institutions are caused by the stability of the relationships, which in turn are safe-guarded by sanctions imposed by members of the community on norm violators (collective disrespect and/or exclusion of norm violating agents). The commitment of agents to specific norms of a group can have several reasons: the interest of an agent to be member of a trustworthy network (i.e., social capital), to be acknowledged for honorable behavior (i.e., symbolic capital, reputation), while norm violation would lead to a loss of those kinds of capital. Moreover, commitments also may be due to the adoption of a norm as a certain value of its own or as an unreflected disposition of action (e.g., routine, habit). The advantage of this type of institution is that undesirable actions and strategies of agents can be sanctioned. This refers to both: (1) clearly undesirable actions like fraud and (2) actions which are undesirable, but only problematic, if they occur on a large scale (bad quality, dumping prices). However, sanctions do not enforce norm conform behavior completely, but effect a loss of reputation and lead to exclusion of agents from the interaction in the community. Hence, the autonomy of agents is only affected partially, since they still may act norm violating.

### *2.2.3. Regulative institutions (RI)*

With this type, we refer to codified, formal law that has been brought up intentionally by the legislative in order to regulate certain social facts. Those law-like rules clearly indicate which actions are allowed or forbidden, and what are the possible consequences in the case of violation. The socio-spatial scope spans the whole legal room, i.e. the entire system. The transintentionality and externality of those rules are caused by procedures of legislation, jurisdiction and execution, while their validity can be enforced by sanctions. If the commitment of agents to law is not caused by the adoption (incorporation) of legal prohibitions and commandments as values of their own, sanctions in form of penalties create incentives to act rule conform. The advantage of this type of institution is that those undesirable actions and strategies of agents that (1) are very harmful (deception), (2) exceed a certain level of occurrence, so that they can not longer be prevented or stopped by normative sanctions (e.g., price dumping), or (3) cannot be resolved by reputation or moral disrespect at all (e.g., monopolies) are regulated by law. However, regulative institutions signify a direct intervention in the agent's autonomy, especially in case that a prison sentence is imposed.

## 2.3. Mechanisms of generating, adapting and reproducing institutions

The description of the three types of institutions already provided some insights how institutions emerge and become prevalent, valid and binding. Rules can be generated in different ways. They either can be laid down intentionally by some social entity or they can emerge bottom-up through the repeated interactions between agents. Rules do not necessarily need to be formulated, established intentionally or to be codified to be valid. But in contrast, even if they are established by a single act (like laws), they need to be adopted and accepted by the agents and reproduced through their actions in order to be valid. According to Bourdieu (1990: 76 pp.), five general mechanisms involved in the process of institutionalization of any type of rule can be distinguished, while peculiar modes can be specified for each type (cf. Table 1):

- The generation and reproduction of rules as well as the change of an institutional type depends to a large extent on the mode of *reflection* about both the rule itself and the obedience to it. While the generation of practical institutions (PIs) does not mean that agents either need to intend to generate a collective style consciously or to act accordingly, the generation of normative institutions (NIs) necessitates discourses and reflections about which behaviors should be valuated in which way. Such discourses may happen in case that agents become aware that a former practical institution has become problematic. If legislative actors or instances discover that some behavior that formerly has been ensured by the means of practical or normative institutions becomes problematic, they may intentionally formulate a regulative institution (RI). Moreover, the reproduction of NIs and RIs by rule-conform behavior depends to some extent on the anticipation of the consequences of rules conform or violating behavior.
- The type of an institution changes depending on the degree to which a specific pattern of behavior is formulated as a rule. While PIs are hardly communicable, those patterns may exhibit inconsistencies and irregularities. NIs are more formalized and communicable. However, in contrast to law, which is intentionally established, formulated and codified and which is meant to be logically consistent, defining a corpus delicti precisely, NIs are more ambiguous and fuzzy.
- Institutions as sociological macro-phenomena have a certain socio-spatial scope, even though this may differ depending on the type of institution. In order to generate an institution or to change its type, it is necessary that either the assignable rules or the class-appropriate behaviors and strategies are diffused within a particular social space. While laws are made available to the entire population by publication, NIs are often spread and generated by gossip and denunciation of dishonorable behavior within a network. PIs itself are not communicable, however, agents of a certain class can comment the behavior of others regarding its feasibility and they may observe and imitate the behavior of others agents of their own class.
- In order to be durable, any institution needs to be objectified, i.e., to become an objective fact, external to the will of agents. In the case of PIs, this is mainly achieved by the incorporation of the collective style into the dispositions of agents (their habitus or beliefs), and by the transformation of a collective behavior into a taken-for-granted certainty. RIs are additionally objectified by forms of materialization, i.e. they are backed by material resources like courts, police etc. Also NIs can be objectified by material resources, e.g., certain monitoring associations.
- The reproduction of a rule by conforming actions of agents, as well as the change of an institution into a type that restricts the autonomy of agents to a larger extent depends on the acceptance of that rule as legitimate. While PIs are perceived as legitimate by the agents of a class as long as they provide feasible strategies of action and serve the purposes of those agents, this legitimating reason is problematic with respect to NIs and RIs, since those institutions often try to prevent behavior that is rational from the perspective of single agents, but not from the perspective of the entire agent population. Therefore, those institutions are often legitimated by their contribution to common goods, welfare and public interests.

### 2.4. Sources of institutional dynamics

The question what are the driving forces that generate, reproduce and adapt institutions still remains, since institutions themselves are not capable of acting. While RIs are established intentionally by agents, the other types of institutions are somehow induced less intendedly by agents who pursue their interests. Moreover, any type of institution needs to be reproduced by agents. However, this does not mean that only individuals are the driving forces of institutionalization processes. Also collective and corporative agents (i.e. groups or organizations that appear as single agents through their representation towards their social environment) can start institutionalization processes. We call those driving forces of institutionalization processes that are not individual agents and that pursue other interest than economically thinking provider agents *instances*. With respect to our application scenario, we distinguish the following instances that generate and safe-guard the different institutional types:

- *Instances of diffusion*: In the economic field, consultants as well as specialist journals and newspapers contribute to the diffusion of feasible strategies (behavioral patterns or plans), valuations of honorable, morally approved behavior, and information about the current state of the market.
- *Reputation networks*: Another instance to diffuse valuations of honorable behavior are reputation networks, in which gossip is spread. These networks also improve the process of building models about competing agents in the market.
- *Associations*: In markets, provider associations often play an important role in establishing NIs with respect to dishonorable providers and unfair competition (e.g., price dumping). Since those collective actors are often accredited, they are powerful regarding sanctions by communicating dishonorable behavior of certain agents. Moreover, they can sanction member agents by excluding them from the association in case of rule violation. An association provides secure trading conditions because of the trustworthiness of the trading partners.
- *State*: With respect to RIs, the different corporate actors of the state, i.e. the relevant instances of legislation, jurisdiction and the executive, generate and safe-guard laws.

## 3. Towards the integration of *institutions* into BDI architectures

As we have argued before, social phenomena like rules and obligations help to coordinate agents' interactions and thus to improve the agent's performance (cf. Tennenholtz, 1998), the regulation of e-commerce (cf. Dignum, 1999) and open electronic marketplaces (cf. Dellarocas, 2001). Moreover, we agree with Lopez y Lopez et al. (2004): "(A)gents with the ability to autonomously determine which norms to fulfill, and which societies to be a part of, are clearly necessary if we wish computational entities to automatically create open societies with others".

Although the improvement of the agent's performance sounds very promising, complying with institutional rules and obligations is not always the most rational way to fulfill an agent's goal as (1) institutionalized rules may directly conflict with the

**Table 1** Primary mechanisms of institutionalization and sources of institutional practice (instances)

| | Practical institution | Normative institution | Regulative institution |
|---|---|---|---|
| **Mechanisms** | | | |
| Reflection | Generation: implicit<br>Obedience: prereflexive | Generation: discursive<br>Obedience: (pre)reflexive | Generation: intentional<br>Obedience: reflexive |
| Formalization | — | Formulation | Codification |
| Officialization | Comments about actions | Communication of rule<br>Valuation of actions | Publication<br>Claim |
| Objectivation | Incorporation<br>Self-evidence | Materialization<br>Naturalization | Materialization |
| Legitimation | Feasibility<br>Conclusiveness | Common good<br>Morality | Public interest<br>Public welfare |
| **Instances** | | | |
| Generating instances | Reputation networks<br>Instances of diffusion | Reputation networks<br>Instances of diffusion<br>Associations | Legislative |
| Safe-guarding instances | —<br>— | Reputation networks<br>Associations | Judiciary<br>Executive authority |

agent's desire or intention or (2) the compliance of two or more rules is impossible, since the action needed to fulfill a single rule violates other institutions that could have been established on the market. On the other hand, intelligent violation of particular institutions can be profitable. Consequently, an agent may have to rationally decide (1) whether to comply with institutional rules, (2) whether to adapt institutional rules that become then obligatory and (3) whether to strategically violate active institutions.

As we have argued in Section 2, rule compliance can firstly be carried out implicitly and pre-reflexive, based on the agent's beliefs. Secondly, they can be a consequence of reflections or reasoning about negative incentives (sanctions). But note that in case of sanction-based institutions, the compliance with them can also provide additional utility, i.e., incentives like reputation for honorable behavior (symbolic capital) or advantages due to the affiliation in a network (social capital). Moreover, agents may attribute a value to some rules for its own sake and incorporate this rule into their own dispositions (beliefs and intentions). As a consequence, we should provide two modes to select actions: (1) institution-reflecting pursue of interests (desires, goals) and (2) disposition-based rule compliance in the sense that those rules are taken for granted and no alternative (rule violation) is taken into consideration. In contrast, institution-reflecting actions require the explicit representation of a rule in the internal structure of the agent. Hence, in order to allow different attitudes towards institutions, we have to provide some basic functionality. Firstly, the agents should be able to recognize the existence of rules (NIs, RIs) on the one hand and to adopt certain behavioral patterns (PIs) or rules (NIs, RIs) into their beliefs, desires and intentions on the other hand. Secondly, the agents need to be provided with a kind of reasoning mechanism allowing them to decide whether certain dispositions or rules should be adopted, or if rules should be deliberatively followed or violated (in case of NIs and RIs). Thirdly, if the agent adopted a rule, it should be able to react on deviant behavior by imposing sanctions corresponding to that institution.

In many applications in the multiagent field, norms are treated as built-in constraints, which is, from our point of view, crucial for open MAS. In contrast, we prefer an approach where institutions emergence during runtime. With respect to e-markets, the most common architecture is described by Bratman (1987) that bases on *beliefs*, *desires* and *intentions*. However, social concepts like institutions, i.e., obligations and rules are not considered, which has prompted us to extend this framework by adding the component *institutions*:

- The agent's *beliefs (B)* represent the mental model of the world (the market) including the consequences of its action that bases on personal perception and may differ from agent to agent. Regarding electronic markets, these beliefs include the agent's current knowledge about the structure of the economic field (competitive positions of other providers), the demand situation and its own position as well as knowledge about achievable profits (economic, cultural, social, and symbolic capital) illustrating the agent's performance on the market.
- The agent's *desires (D)* reflect a projection that the agent wants to pursue. Regarding e-markets, desires can vary with respect to the sort of capital an agent wants to accumulate.
- The agent's *intentions (I)* translate a goal into actions. They are future directed and often lead to the execution of a plan. In our scenario, these intentions are feasible competitive strategies, consisting of a combination of possible courses of a number of actions (plans). Committed plans may modify the belief sets, activate institutions or create new desires.
- The set of *institutions (Inst)* reflects the current institutional rules and possible sanctions related to that institution an agent has access to. The influence of institutional obligations and information about practical strategies on an agent's reasoning process may differ from agent to agent.

In the following, we discuss (1) how institutions may influence the agent's reasoning process, (2) under which conditions an agent adopts an institution and (3) how conflicts between the BDI components are resolved.

### 3.1. A meta-level reasoning process

The selection of the most suitable plans to reach a set of goals is the result of the agent's internal deliberation process, where a plan defines the priorities in keeping/achieving certain levels of capital (cf. Bourdieu, 1998) and can again be decomposed into several sub-plans that can be distinguished by actions that illustrate the different options the agent has to fulfil the sub-plan (cf. Table 2). Thereby, a sub-plan is not just a sequence of basic actions, but may also include more abstract elements such as sub-goals. When an agent decides on pursuing a goal with a certain plan, it commits itself (momentarily) to this kind of goal accomplishment and hence has established a so-called intention towards the sequence of plan actions.

At the begin of each round, the provider agent updates the key data describing the performance and adequacy of its plans and thus the success in reaching its goal. These key data include (1) the agent's business volume, costs thereby incurred and produced profit, (2) the number of proposals the agent receives (i.e., the percentage of assigned tasks with respect to the agent's cultural capital to measure the workload) to

**Table 2** Sub-plans and corresponding options for actions

| Sub-plans | Options | | |
|---|---|---|---|
| Self-organisation | Create organisation | Resolve organisation | Extend organisation |
| Organisation | Change organisational structure | | |
| Contracts | Comply with contracts | Offend against contract | Offend sometimes against contract |
| Capacities | Buy additional capacities | | |
| New products | Invent new products | | |
| Quality | Improve quality of products | | |
| Request of reputation values | Request reputation from neighbours | Request reputation from organisational members | Request reputation from journalists |
| Diffusion of reputation values | Propagate correct reputation | Propagate 50% false reputation | Propagate false reputation |
| Associations | Create association | Leave association | Violate associational norm |
| Handling of rules | Compliance with rules | Compliance with rules, depending on situation | Deviant behaviour |
| Handling of norms | Compliance with norms | Compliance with norms, depending on situation | Deviant behaviour |

estimate the actual supply-demand configuration, (3) the behavior of the competitors (e.g. cheating or deviant attitude) and (4) any type of institution that have been newly activated.

According to these beliefs and the current agent's intentions, new desires are formed. Additionally, at this stage, institutional rules are incorporated into the set of desires, as system states resulting of active institutions would be preferred as the benefits of acting according to institutions are considered to be higher compared to the current environmental situation. This preference is recognized as a result of (1) increasing cases of fraud, or (2) reduced percentages of assigned tasks due to a lower level of symbolic capital that could be compensated by acting institutional-conform. As a consequence, institutions and the reasoning about the adaptation of institutional behavioral patterns have an impact on the agent's goal formation process. If an adequate institution has not been activated yet, the agent can take this occasion to activate an adequate type of institution. The adequacy of an institutional type is discussed in Section 3.2.

In a second stage, the most important goals are transformed into intentions, where the importance of some goal is finally evaluated with respect to the *usefulness* of those plans that could theoretically achieve the goal. In principle, the usefulness of a plan is expressed by a combination of all four sort of capital (cf. Bourdieu, 1998), where capital denotes any kind of '*resource*' that confers status and power to an agent:

- *Economic capital* exists either in institutionalized form as property rights or in materialized form and is directly convertible into money.
- *Cultural capital* can be accumulated in the form of incorporated skills, in materialized form as cultural goods, or in an institutionalized state (guarantee of an accredited third party that certain abilities are really existent, e.g. by certificates or diplomas).

- *Social capital* derives from the membership of an agent in "durable networks of more or less institutionalized relationships of mutual acquaintanceships and recognition" (Bourdieu 1998: 51). The individual possession of this form of capital depends on the size of the network and on the volume and structure of capital that the members possess.
- *Symbolic capital* derives from certain attributes that are collectively ascribed to an agent or collectivity (group, organization, network). These attributes derive to a large extent from the other sorts of capital which an agent possesses, because any kind of capital additionally "tends [...] to function as symbolic capital" (Bourdieu, 2000: 242) if it is recognized as scarce and exceptional.

The combination of all four capital sorts describes the overall performance of the agent's behavior in the last round(s) and includes the profit an agent makes in the last round, the percentage of assigned tasks with respect to the agent's overall cultural capital and the percentage of assigned tasks of the organizations, the agent is a member in, with respect to the organizations' overall capacities. The components of the usefulness functions are equally weighted to ensure that each capital sort is similarly considered in the plan selection process. To weight the capital combination in the favor of economic capital, would result in a more selfish agent's behavior etc. Furthermore, on the basis of the agent's experiences, the usefulness of a plan considers not only the capital accumulation in the next round $t$, but tries to estimate the effects on the performance when applying on the further rounds $t + 1$ (e.g., the influence of malicious behavior on the performance criterion *percentage* of *assigned* tasks in the near future).

Beside, the evaluation of a plan's expected usefulness, and thus the specification of a plan's postconditions, preconditions have to be determined to express a plan's environmental adequacy as a plan's usefulness may differ when applying on different environmental circumstances.

Therefore, we classify different environmental states with respect to the agent's beliefs in terms of performance and market structure and store each plan and its usefulness (e.g., the minimum, maximum and average) that have been applied on the environment The environmental plan repository is extended if new plans are applied on the market and provides a set of feasible behavioral patterns describing how to act and behave with respect to the environmental states (i.e., preconditions). If the agent recognizes the existence of practical institutions and those behavioral patterns encourage to arrive the set of goals, the corresponding institutional rules are incorporated and the agent's plan is adapted.

After evaluating the plan that shows the highest usefulness, possibly new goals have to be generated as for instance the non-compliance of particular institutions that are not obliged for the agent would reduce the plan's usefulness and thus the agent's performance (e.g. negative side effects) with respect to its goals. Thus, we agree with Castelfranchi et al. (1999) that institutions impact the goal generation process, although these goals are not directly desired by the agent. Especially for normative institutions, agents do not have to fear any sanction on monetary basis if they do not act according to the institutionalized behavior, but if they interact with normative agents, the probability of getting further task announcements in the near future decreases due to their deviant behavior which could complicate the achievement of other goals like the improvement of economic or symbolic capital and would thus result in conflicting goals.

Finally, the selected goals are transformed into intentions and the corresponding plans are executed. If the executed actions fail, the planning process is restarted on the basis of this new information by changing the corresponding sub-plan(s). Due to limited information, the main problem consists in identifying the most adequate plan to fulfill the agent's goals. This lack of knowledge may lead to unfeasible or inadequate strategies that could lead to poor performance and unacceptable system states. Initially, the consequences and effects of plans are not known by the agents, instead, the classification of a strategy's adequacy with respect to environmental and internal circumstances must be learned by applying those plans on the market. The evaluation of the corresponding usefulness gives then information about the profitableness of certain strategies to reach a set of goals. After collecting enough experiences, this classification allows that providers flexibly adapt on different environmental conditions and implies the formation of a case-based reasoning system where the environmental states operate as classification attribute.

Similar to our approach, Broersen et al. (2001) developed the BOID agent architecture that uses obligations as additional attitudes besides beliefs, desires and intentions. Agents have to resolve conflicts between their attitudes (e.g., between goals and intentions), where conflict resolution types determine orders how to overrule between the attitudes. Using this architecture, agent's characteristics like being selfish or stable can be defined by specifying which component overrides the others. In contrast to BOID, NoA (cf. Kollingbaum and Norman, 2003) agents can reason about how conflicting norms are resolved. Of particular interest in this approach is the level of consistency that occurs if an agent adopts a new norm. Three levels of consistency are described: strong consistency, weak consistency and strong inconsistency. Using the classification into levels of consistency, an agent may make reasonable decisions on whether or not to adopt a new norm.

Lopez y, Lopez, Luck and d'Inverno (2002) present a proposal for agents that make a decision whether or not to adopt a norm, which includes issues such as the attitude of the agent towards norms, or the consistency of a norm with the agent's currently adopted goals.

In contrast to these approaches, the agents in our framework resolve conflicts between the BDI components by ranking the usefulness (i.e., adequacy) of the plans that could achieve the agent's goal for the current environmental states. Consequently, the agent's characteristics with respect to the compliance with institutions depend on the agent's welfare and performance in the e-market. The developed types of institutions do not stretch different levels of consistency. Instead, the institutional rule that shows the highest usefulness is adopted, regardless of the type of institution.

### 3.2. The activation of institutional types

In principle, we take up the argument of Conte and Castelfranchi (2001) in our arrangement of the institutional types. They define social order as a pattern of interactions among interfering agents that allows the satisfaction of the interests of some agents. The interests of particular agents become apparent through activated and existing institutions. The foundation that institutions emerge depends on (1) an agent's intention to activate an institution and (2) the agreement of at least a part of the society which also intents to adopt this institution. In the last section, we have seen how and when *institutional* goals are formed. In this section, we illustrate under which conditions the different types of institutions are activated.

### *3.2.1. Activation of practical institutions*

Practical strategies are not consciously activated. Only if the intentions of some agent demonstrate usefulness in reaching a particular goal, journalist agents diffuse the corresponding sub-plan mechanism. A practical institution is finally activated if the intentions of this agent are reproduced and adopted by any other agent of the society, as its developed plans have not proven to be profitable and practical.

### *3.2.2. Activation of normative institutions*

Normative institutions can be activated, if at least a sub-population has the goal to establish explicit norms and obligations. All agents agreeing in certain norms are obliged to act norm-conform. As in the case of regulative institutions, the source of this type of institutionalisation is that some agent has the intention to establish norms. The motive to come to this decision may be (1) the spread of gossip, (2) breach of contracts during the interaction etc. These occurrences may negatively affect the economised capital and the agent's reputation influencing the probability of getting tasks announced in the next rounds.

### *3.2.3. Activation of regulative institutions*

In general, regulative institutions are activated by the agent society, if the system-designer has specified this kind of institutional rule as honourable behaviour at design-time. Indeed, the impulse itself to establish a regulative institution must be initialised by some agent having the interest to establish explicit rules. For this purpose, the preconditions are that (1) the agent forms an institutional goal intending the activation of obligations in the agent society and (2) that normative institutions are no solution to handle this form of perturbation.

## 4. A preliminary report from simulation studies

This model of flexible self-regulation by institutions has been implemented as a multiagent-based market.[2] In order to develop a model of the market itself, we have chosen the market of transportation and logistics as application scenario. The goods to be exchanged by customer and provider agents represent transportation services or, more abstractly, 'tasks' that can be of different types (A, B, C, etc.), different capacities (A, AA, AAA, etc.), different qualities (A, A+) and have a deadline (latest delivery time). The current version of our market model consists of three different types of

[2] The simulation software was implemented in JAVA. The software, both source and executable, is available at: www.ags.uni-sb.de/∼chahn/JTIM.

agents that differ with regard to their architecture: while (1) *providers* are equipped with the BDI-architecture described above, (2) *customers* are modeled as reactive agents that fix their demand any five rounds with regard to size (capacities), types, and qualities according to the current average supply of the provider population. Moreover, after sending an offer and receiving a proposal in turn, customers assign tasks to providers on the basis of four criteria (reputation, quality, price, and social capital of the respective provider). (3) *Journalist agents*, as instances of diffusion, interview providers and customers in order to receive '*reputation values*' of a target agentwith regard to its *credibility* (correctness of its reputation statements), its *trustworthiness* (compliance of e-contracts, i.e., task fulfillment) and its *skills* (tasks, qualities, capital assets). On the basis of this information, journalists provide '*reputation reports*' by sorting out deviating information of apparently lying witnesses,selling them on inquiry to customers and providers,and *market analyses* that they sell to subscribers.

For this paper, we restricted the empirical evaluation of our model by simulation studies to the perturbation scenario of *deviant provider agents*. Deviance is defined by two behaviors: while '*fraud*' refers to breaches of e-contracts, '*deception*' occurs if agents exchange reputation values of others and lie, spreading false information intentionally. Decisions regarding fraud and deception depend on the reasoning process of the providers, i.e., which option of the sub-plans 'diffusion of reputation values' and 'contracts' they consider as appropriate strategy (cf. Table 2). 200 simulations have been run, each for 50 rounds, starting with 20% of the provider population executing the sub-plan 'offend against contract'. During the simulation, any provider is autonomous to abandon or adopt this sub-plan.

In order to investigate the *impacts* of the institutional types and the corresponding instances on the system's capacity for self-regulation *separately* before evaluating the entire self-regulation process, we firstly restricted the simulation study to the types '*practical institution*' and'*normative institution*'. The underlying question has been whether the rule *not to deviate* (i.e., not to break e-contracts or to spread lies) can be institutionalized as a 'style of action' or if the rule requires to be institutionalized as 'social norm' and sanctions in form of moral disrespect (denunciation of deviant behavior leading to a loss of symbolic capital) are necessary. Secondly, we restricted the study to the investigation to three different types of institution generating instances (cf. Table 1): in 50 of the 200 runs either '*reputation networks between single agents*', '*reputation networks between organization members*', '*journalists*', or '*no instance*' are allowed.

'No instance' means that agents are only permitted to calculate the values for trustworthiness and skill of other agents on the basis of their own experiences. In simulations with 'reputation network between single agents', agents are additionally allowed to request values for credibility, trustworthiness, and skill of other agents. 'Reputation networks between organization members' differ with regard to the preferences of providers whom to ask for reputation values. Members of the same organization are preferred since the relation between providers (social capital) has an influence on their decision whether to lie or not. In simulations with 'journalists' as instances of diffusion, agents are allowed to request reputation values from 5 journalist agents in addition to their own experience, but not to ask others. All runs start with a provider population of 50 agents. Initially, any provider is equipped with the capacity of one 'A', but may expand its capacities or become insolvent during runtime. The number of

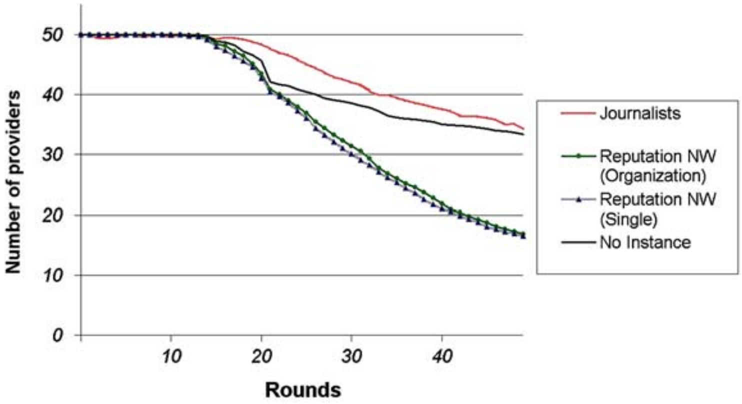

**Fig. 1** Number of provider agents

customers constantly remains at 20. Initially, any customer demands a single ‘A’. This means that at the start of each run, only 20 of 50 providers receive an acceptance of their proposal so that the factual number of cheaters in the first round varies randomly, depending on how many of the 20% initial cheaters actually get a task assigned. After five rounds the demand is adjusted to the factual supply in the system as described above.

Each of the four configurations has been investigated with respect to different *performance measures* (economic stability, reliability of task assignment and fulfillment, reliability of information diffusion) in order to determine the contribution of each instance and the respective mechanisms to the institutionalization of non-deviating behavior and moreover, to the robustness of the system. A general result of the simulations is that the different instances do not have the expected impacts with respect to stability and reliability (cf. also Hahn et al., 2005). We hypothesized that (1) configurations that permit the instance ‘reputation network (single)’, ‘reputation network (organization)’, and ‘journalists’ have different impacts on robustness, but perform better than ‘no instance’. Moreover, we assumed that (2) the larger the scope of an instance regarding the diffusion of reputation values the better the performance of the respective configuration.

In fact, the latter hypothesis is not falsified by the results regarding *economic stability*, measured by both the overall as well as the individual accumulation of economic capital. Simulations using ‘journalists’ perform very much better than those using ‘reputation network (organization)’ with an overall accumulation of 2102 units economic capital on system-level in round 50 for ‘journalists’ in comparison to 182 units for organizational reputation networks. Furthermore, organizational reputation networks perform slightly better than single agents’ reputation networks that achieve only 134 units. Nevertheless, the highest average overall economic capital is accumulated by the configuration that does not permit the officialization of deviant behavior (2947 units). These results on system-level regarding stability correspond mainly with the results on the agent-level (cf. Fig. 1).

For all simulations in which information is diffused by instances, the average number of agents declines continuously after round 13 because of insolvencies. ‘Reputation

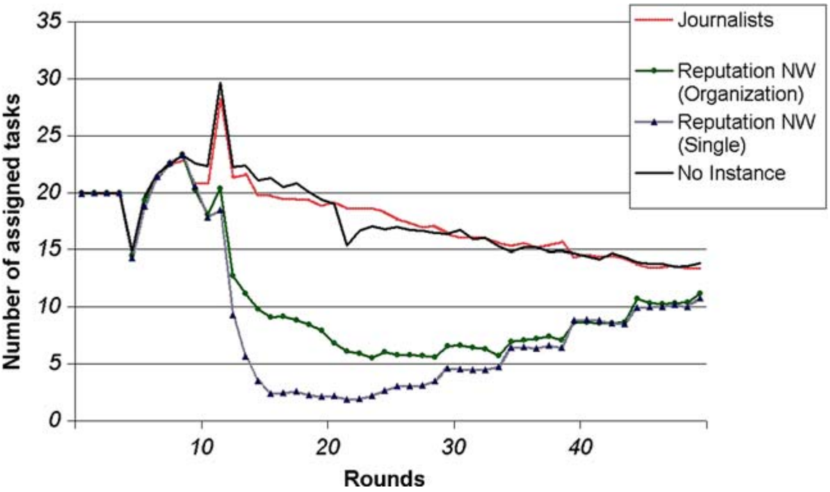

**Fig. 2** Number of assigned tasks

network (single)' performs worst (ending up with 17 agents). 'Reputation network (organization)' realizes only little better results (also 17 agents at the end). 'No instance' (33) and 'journalists' (34) definitely enable more agents to survive economically. In contrast to the accumulation of economic capital on system-level, 'journalists' actually perform better than 'no instance' in terms of maintaining a high number of agents.

This bad economic performance of configurations that allow the sanctioning of deviant behavior by diffusing reputation values in comparison to the configuration 'no instance' apparently has several reasons:

The primary reason is the performance in terms of the system's *reliability concerning task-assignment* (cf. Fig. 2). Before round 12, the amplitudes of the graphs that represent the number of assigned tasks (delivered tasks and non-delivered tasks due to fraud) are mainly caused by the adjustment of the customers' demand after round 5 and 10. The number of assigned tasks declines in round 5 since a new demand of products requires that providers collaborate. An analysis of the number of agents being members of organizations showed that agents successfully form organizations between round 5 and 10 so that the number of assigned tasks reaches a climax for 'reputation network (organization)' and 'reputation network (single)' in round 9 and for 'journalists and 'no instance' in round 12. Apparently, the customers augment their demand again in round 10.

However, after round 12, the agent populations for all reputation types including 'no reputation' are not able to trigger a new increase of demand. Partially, this development is caused by first cases of insolvency (cf. Fig. 1) in round 12 that diminish the agent population and hence the capacities of the system so that a vicious circle starts: each decline of demand due to a decrease of the agent population influences the economic performance of the remaining agents negatively, leading to a new shrinkage of the agent population and vice versa. However, this development is not completely endogenous since the graphs for task-assignment and number of agents vary considerably between the different simulation configurations. While the number of assigned tasks and the

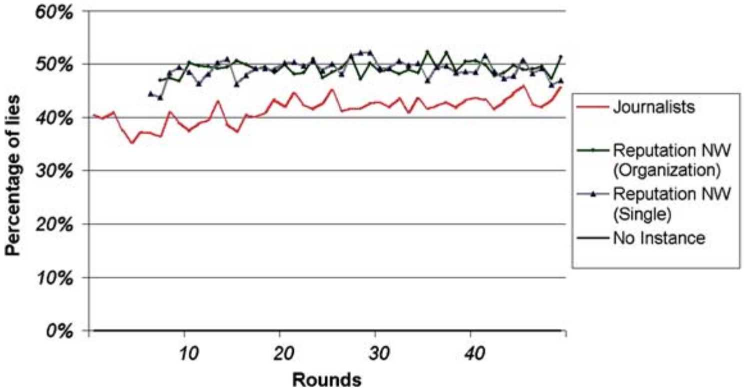

**Fig. 3** Percentage of lies

number of agents for both ‘image’ and ‘social esteem’ fall heavily after round 12, but recover (except the number of agents) later again, the graphs for ‘no reputation’ and ‘prestige’ nearly decline monotonously with a lower negative gradient.

One central reason of these differences is the system’s *reliability concerning information diffusion* measured by the rate of lies (cf. Fig. 3). After agents start to exchange reputation values in round 7, they intentionally spread false information at a rate around 50% in organizational reputation networks as well as among single agents’ reputation networks. If journalists collect information, providers lie at a rate of 40%. The lower rate is most likely caused by the influence of lies on the economic performance because providers receive a payment for giving interviews, but are only interviewed again if their information does not deviate continuously from the average and expose them as liars.

Nevertheless, the high rate of lies for all types of instances and the insignificant influence of reputation on the development of this rate is an open research question at present because bad values for ‘credibility’ lead to a loss of symbolic capital and therefore have been assumed at design-time to ‘convince’ agents to change the selected sub-plan for ‘diffusion of reputation values’. At least, the high rate of false information explains the generally bad, but differing performance of the configurations that allow to spread reputation values: the higher the number of lies the lower is the capability of both customers and providers to find appropriate exchange or cooperation partners because only little reliable information about the skills and the trustworthiness of providers is available. But while the high rate of deception leads to an ruinous effect if lies are spread in reputation networks, an extreme decline of performance can be prevented in case of journalists that evaluate the correctness of received information. But still, the influence of the high rate of deception lowers the performance in contrast to simulations without any information diffusion.

The high rate of false reputation is also a reason for the generally insufficient impact of information diffusion on the *reliability regarding task fulfillment* (cf. Fig. 4) and, consequently, a cause of the bad performance concerning task-assignment and economic stability. While the initial fraud rate, preconfigured by the designer, firstly drops rapidly without noticeable differences between the four

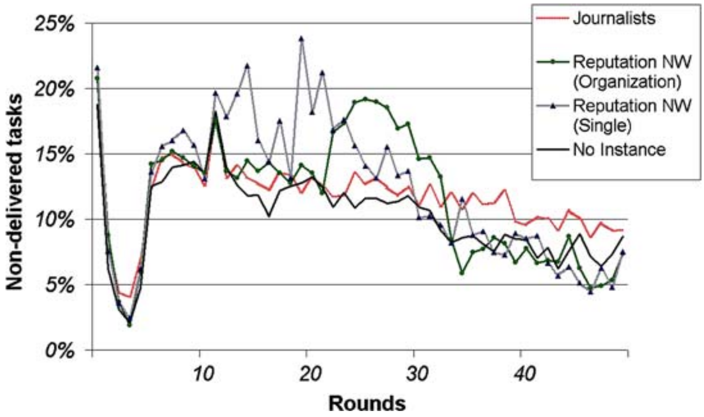

**Fig. 4** Percentage of fraud

configurations, it raises again after round 5. Nevertheless, for 'journalists' and 'no instance' the percentage of non-delivered tasks does not exceed the initial level again and finally declines to 9%. Consequently, the economic losses produced by fraud can be reduced slowly during runtime, but incur constantly, additionally explaining the negative gradient of the graphs for task-assignment and number of agents for these two types. In contrast, the rate of fraud augments for the two types of reputation networks much more after round 12, exceeding the initial 20% in some rounds. For both types, it takes much longer until they decrease again. However, finally the rates for both types are lower than for 'journalists' and 'no instance'.

The different development of the rate of deception and the rate of fraud raises the question whether these differences are caused by *external interventions into the agents' autonomy to cheat*, namely through the exclusion of cheaters from trade taking providers the chance to select the plan 'offend' against e-contracts' again. Or is the decline of fraud rates a *result of reasoning processes*, produced by changes of cheaters' beliefs and internal states either due to experience or to information learned by others?

Since fraud rates decline for all of the four configurations, no matter whether the spread of reputation values is permitted or not, we assumed that the reduction of fraud is rather caused by experience (providers learn in which cases fraud is profitable) than by information diffusion. Thus, to evaluate the influence of the BDI-process for this paper, we analyzed the development of the ratio between agents that already broke their e-contract in the first round and all cheaters per round (cf. Fig. 5). Although the rates for all configurations show a similar pattern (the graphs slightly increase until the last round) and nearly proceed at a same level (around 50%), the rate for 'no instance' is the lowest, while the rates for both reputation networks grow most intensely. This allows the conclusion that spreading reputation values apparently *does not further the exclusion of cheating agents*. The ratio between early cheaters and agents starting to break contracts in later rounds cannot be reduced more in simulations with information diffusion than in those without any information exchange.

The result, that the decline of the fraud rate must be caused by the reasoning of agents and their experience under which conditions fraud has positive or negative consequences regarding capital accumulation, is also confirmed by comparing the

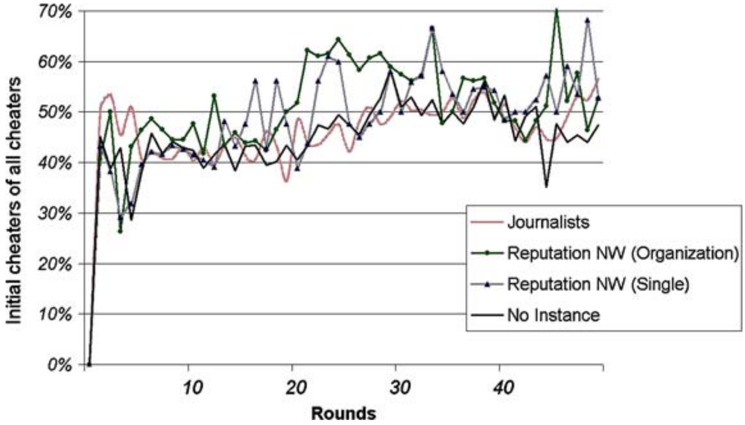

**Fig. 5** Percentage of initial cheaters (all simulation runs)

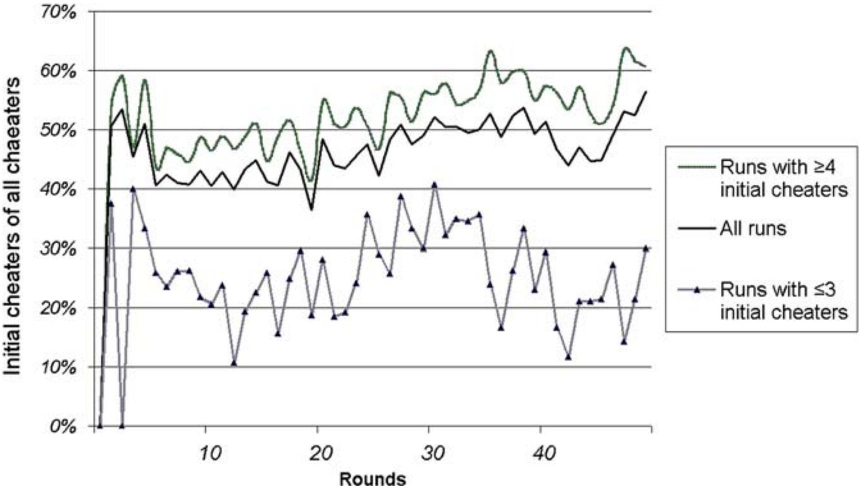

**Fig. 6** Percentage of initial cheaters (journalists)

ratio of initial cheaters and late cheaters per round between different simulations runs of the same configuration. As mentioned above, although the initial fraud rate averages for all runs at a level of 20%, the factual number of initial cheaters varies randomly for each run because not all 50 agents including the 20% of preconfigured cheaters get a task assigned in the first five rounds. So the number of cheaters that make the experience that fraud pays because they received a task varies. In order to investigate whether the number of initial cheaters influences the later development of the probability that these agents cheat again, we split the simulation runs into (1) runs with a number of three and less and (2) with a number of four and more initial cheaters. Figure 6 for the configuration 'journalists' shows that indeed, the ratio of initial cheaters remains higher during runtime if the number of initial cheaters has been higher than three.

An analysis of the other configurations as well as of the absolute numbers produced similar results.

## 5. Conclusions and future work

In this paper, we argued that allowing agents autonomously to activate rules (social institutions) and to regulate the type of institutions during runtime has beneficial effects on the robustness of open electronic marketplaces. Referring to sociology, we presented a notion of social institutions consisting of more than pure rule systems. Reputation networks, associations, journalists and an executive power (summarized as 'instances') as well as norm-autonomous single agents are considered to be necessary prerequisites to ensure the activation and compliance of rules. We argued that especially instances influence the behavior and performance of agents and hence the entire system in different ways. Consequently, each of the three institutional types is only to a certain degree appropriate to regulate the different perturbations scenarios sketched in the introduction since the types affect the agents' autonomy differently. However, first simulation studies on the perturbation scenario of deviant agents showed that certain institutional types (practical and normative institutions) and the respective instances and mechanisms of institutionalization (reputation networks and journalists) have no significant beneficial impact on the system's robustness compared to simulations that do not permit agents to spread information about deviant behavior. Nevertheless, in order to disprove our initial hypotheses, more simulations are needed to investigate whether other institutional forms or instances are more appropriate with respect to this perturbation scenario. Additionally, studies on simulations enabling the entire self-regulation process and allowing agents to activate institutional types autonomously during runtime are required. Finally, other perturbation scenarios need to be simulated.

**Acknowledgment** This work was funded by the German Research Foundation (DFG) in the priority program Socionics under contracts FI 420/4-3 and FL 336/1-3. We are indebted to Dr. Klaus Fischer, Dr. Frank Hillebrandt, Daniela Spresny and Ingo Zinnikus for their constructive ideas and comments, and to Rolf Schmitt for his support concerning implementation and simulations. A previous version of this paper has been presented at the First International Symposium on Normative Multiagent Systems (NorMAS 2005) at the AISB Convention in Hatfield, Hertfordshire, England, April 2005.

## References

Axtell R (2001) Effects of interaction topology and activation regime in several multi-agent systems. In: Proceedings of the 2nd International Workshop on Multi-Agent Based Simulations, Boston, MA USA, July 2000, Lecture Notes in Artificial Intelligence (Vol. 1979), Springer, Berlin Heidelberg New York pp 33–48

Berger PL, Luckmann T (1966) The social construction of reality. Doubleday, New York

Boella G, Damiano R (2002) A game-theoretic model of third-party agents for enforcing obligations in transactions. In: Sartor G, Cevenini C (eds) Proceedings of the Workshop on the Law of Electronic Agents (LEA02)

Bourdieu P (1990) In other words. Essays towards a reflexive sociology. University Press, Polity Press, Stanford, Cal., Cambridge/UK

Bourdieu P (1998) The forms of capital. In: Halsey AH, Lauder H, Brown P, Stuart Wells A (eds) Education. culture, economy, and society. Oxford University Press, Oxford, New York, pp 46–58

Bourdieu P (2000) Pascalian meditations. Stanford University Press, Stanford/Ca

Bratman M (1987) Intentions, plans and practical reason. Harvard University Press, Cambridge, MA

Broersen J, Dastani M, Hulstijn J, Huang Z, van der Torre L (2001) The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In: Proceedings of Autonomous Agents 2001, pp 9–16

Castelfranchi C (1999) Prescribed mental attitudes in goal-adoption and norm-adoption. Artificial Intelligence and Law 7(1):37–50

Castelfranchi C (2000) Engineering social order. In: Omnici A, Tolksdorf R, Zambonelli F (eds) Proceedings of the First International Workshop (ESAW'00) on Engineering Societies in the Agents World, Lecture Notes in Artificial Intelligence 1972, Springer Berlin, Heidelberg: pp 1–18

Colombetti M, Fornara N, Verdicchio M (2002) The role of institutions in multiagent systems. Atti del VII convegno dell Associazione italiana per l intelligenza artificiale (AI*IA 02), Siena

Conte R, Castelfranchi C (2001) Are incentives good enough to achieve (info)social order. In: Dellarocas C, Conte R (eds) Proceedings of the Workshop on Norms and Institutions in MAS (at AGENTS2000), Barcelona, Spain, 2000

Dellarocas C (2001) Negotiated shared context and social control in open multiagent systems. Social Order in Multiagent Systems, Kluwer Academic Publishers, Boston

Dignum F (2001) Agents, markets, institutions and protocols. In: Dignum F, Sierra C (eds) The European AgentLink perspective, Lecture Notes in Computer Science (Vol. 1991), Springer, Berlin Heidelberg New York pp 98–114

Dignum F (1999) Autonomous agents with norms. Artificial Intelligence and Law 7(1):69–79

Dignum F (2004) Abstract norms and electronic institutions. In: Lindemann G, Moldt D, Paolucci M, Yu B (eds) Proceedings of regulated agent-based social systems, Lecture Notes in Artificial Intelligence (Vol. 2934), Springer, Berlin Heidelberg New York, pp 93–103

DiMaggio P (1997) Culture and cognition. Annual Review of Sociology 23:263–287

Esser H (2000) Soziologie. Spezielle Grundlagen, Bd. 5: Institutionen, Campus, Frankfurt a.M

Esteva M, Rodriguez JA, Sierra C, Garcia P, Arcos JL (2001) Agent-mediated electronic commerce. In: Dignum F, Sierra C (eds) The European AgentLink perspective, Lecture Notes in Computer Science (Vol. 1991), Springer, Berlin Heidelberg New York, pp 126–147

Froomkin AM (1997) The essential role of trusted third parties in electronic commerce, 75 Oregon L. Rev. 49.

Hahn C, Fley B, Florian M (2005) Social reputation: A mechanism for flexible self-regulation of multiagent systems. Journal of Artificial Societies and Social Simulation (under review)

Kollingbaum MJ, Norman TJ (2003) Norm consistency in practical reasoning agents. In: Dastany M, Dix J (eds) PROMAS Workshop on Programming Multiagent Systems

Lopez y, Lopez F, Luck M, d'Inverno M (2002) Constraining autonomy through norms. In: Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-agent Systems, pp 647–681

Lopez y, Lopez F, Luck M, d'Inverno M (2004) Normative agent reasoning in dynamic societies. In: Jennings NR, Sierra C, Sonenberg L, Tambe M (eds) Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2004), New York, USA, 2004. ACM Press, 535–542

Verhagen H (2000) Norm autonomous agents, PhD thesis, Stockholm University

Schillo M, Bürckert H-J, Fischer K, Klusch M (2001) Towards a definition of robustness for market-style open multi-agent systems. In: Proceedings of the 5th International Conference on Autonomous Agents pp 75–76

Schillo M, Fischer K, Fley B, Florian M, Hillebrandt F, Spresny D (2004a) FORM—A sociologically founded framework for designing self-organization of multiagent systems. In: Lindemann G, Moldt D, Paolucci M, Yu B (eds) Regulated agent- based social systems, Lecture Notes in Artificial Intelligence (Vol. 2934), Springer, Berlin Heidelberg New York, pp 156–175

Schillo M, Fischer K, Hillebrandt F, Florian M, Dederichs A (2000) Bounded social rationality: Modelling self-organization and adaption using habitus-field theory. In: Proceedings of the 1st Workshop on Modelling Artificial Societies and Hybrid Organisations (MASHO), pp 112–122

Schillo M, Knabe T, Fischer K (2004b) Autonomy comes at a price: Performance and robustness of multiagent organizations. In: Hillebrandt F, Florian M (eds) Adaption und Lernen in und von Organisationen, Westdeutscher Verlag, Wiesbaden, pp 127–140

Scott RW (2001) Institutions and organizations, 2nd edn. Sage Publications, Thousand Oaks

Tennenholtz M (1989) On stable social laws and qualitative equilibria. Artificial Intelligence 120(1): 1–20

Wooldridge M, Jennings N, Kinny D (1999) A methodology for agent-oriented analysis and design. In: Proceedings of the 3rd International Conference on Autonomous Agents, AA99, ACM Press, New York, pp 69–76

**Christian S. Hahn** studied computer science and economics at Saarland University and received his diploma in 2004. Currently, he works in a project of the priority program ‘Socionics’ funded by the German Research Foundation at the German Research Center for Artificial Intelligence (DFKI).

**Bettina Fley** studied sociology, economics, law, and social and economic history at the University of Hamburg and received her diploma in 2002. She currently works in a project in the priority program ‘Socionics’, which is funded by the German Research Foundation (DFG), at the Department of Technology Assessment at the Hamburg University of Technology.

**Michael Florian**, received his master in sociology at the University of Münster, where he also finished his doctoral degree in 1993. Since 1995, he holds a position as a senior researcher (‘Oberingenieur’) at the Department of Technology Assessment at the Hamburg University of Technology and heads the sociological part of a project in the priority program ‘Socionics’ funded by the German Research Foundation (DFG).