Emotion-based Norm Enforcement and Maintenance in Multi-Agent Systems: Foundations and Petri Net Modeling

Julia Fix
University of Hamburg
Vogt-Kölln-Str. 30
D-22527 Hamburg, Germany
+49-40-42883-2225
fix@informatik.uni-hamburg.de

Christian von Scheve
University of Hamburg
Allende-Platz 1
D-20146 Hamburg, Germany
+49-30-91607790
xscheve@informatik.uni-hamburg.de

Daniel Moldt
University of Hamburg
Vogt-Kölln-Str. 30
D-22527 Hamburg, Germany
+49-40-42883-2247
moldt@informatik.uni-hamburg.de

ABSTRACT
A review of recent theories of emotion indicates close interconnections between emotion and social norms in human societies. We consider the possibility of implementing these mutual influences in a multi-agent system in order to establish dynamic and flexible control structures. According to some theories, emotion plays a key role in establishing and maintaining these structures by fostering the internalization of and compliance with social norms. Here we introduce a Petri Net based approach to modeling the emergence and maintenance of social norms in multi-agent systems.

Categories and Subject Descriptors
I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence - *Intelligent agents*, J.4 [**Computer Applications**]: Social and Behavioral Sciences - *Psychology, Sociology*.

General Terms
Design, Human Factors.

Keywords
Emotions, social norms, MAS, socionics, Petri Net modeling.

1. INTRODUCTION
Controllability of multi-agent systems (MAS) is one of the main challenges for current Distributed Artificial Intelligence (DAI) research: the desire for systems' autonomy, flexibility, and proactivity may conflict with imperative issues of global system control [13]. Achieving system control without abandoning core strengths of DAI systems might be accomplished by using the social mechanisms providing coordination and control in human societies. Applying these mechanisms and their functions to artificial societies is the main goal of this paper.

One feasible solution to the problem just outlined is the implementation of a system of shared norms which is not coerced by the designer, but instead emerges from mutual interactions of agents [15]. To realize such an approach, it might be beneficial to adapt to the computational context the mechanisms of norm maintenance and compliance in human social systems. Proper theorizing and empirical evidence both indicate that emotion is essential in sustaining social norms and enforcing sanctions in cases of non-compliance [6, 7, 9]. This regulatory function of emotion in societal affairs was recently corroborated by empirical findings suggesting that emotions reinforce norm-compliant behavior and punishment of deviant behavior [4, 5].

In support of the plea to integrate theories of emotion into the computational study of social norms [16], we aim to show how emotion theories of different disciplines can jointly specify the interrelation between social norms and emotion. Furthermore, we argue that emotion may provide a basis for establishing solid control mechanisms within MAS by promoting norm-compliant behavior, provided that social norms are at all implemented in the system. First, we offer some initial theoretical foundations and second, we represent selected aspects in a Petri Net based model.

2. THEORETICAL APPROACHES
In this section we will briefly review some social science theories suggesting a crucial function of emotion in sustaining common behavioral norms. Basically, the theories claim that some emotions sustain social norms whereas others (e.g., contempt, disdain, disgust) are incentives to punish violators [6, 8]. "Self-conscious" emotions such as shame, embarrassment, and guilt are, inter alia, evoked by violating a social norm. They are negative payoffs and the desire to avoid them contributes to sustaining social norms within an actor, ancillary to any third party punishment. Thus, the anticipation of certain emotions promotes the internalization and, subsequently, the enforcement and maintenance of social norms.

In addressing the "foundational theoretical problem of the social sciences – the possibility of unconscious, unplanned emergent forms of cooperation, organization and intelligence among intentional, planning agents" [2, p.6], we have argued that the potential of social norms in explaining social structural dynamics cannot be investigated properly without considering the role of emotion [14]. Conte and Castelfranchi view a norm as a mental object, i.e. a hybrid configuration of beliefs and goals [3]. Endorsing this approach, the social, temporal, and spatial distribution of norms makes them instances of a macro system; at the same time, however, their definition as configurations of

AAMAS'06, May 8-12, 2006, Hakodate, Hokkaido, Japan.

beliefs and goals renders them an instance of the micro level and accordingly a primary subject matter of different kinds of reasoning processes. Previously, we have shown that emotion may be a loophole in this respect [14] and that Elster's [8] conceptualization thereof is well suited to serve as an amendment to existing positions.

Elster proposes a definition of certain qualities of social norms instead of the concept as such [8]. Accordingly, norms are "social" insofar as they are sustained through sanctions by a third party which is by definition unaffected by the norm transgression. Consequently, the presence (or representation) of other agents is necessary to enforce and maintain a social norm. Elster argues that imposing sanctions on the norm violator is driven by emotions such as contempt or disgust, whereas "sanction" may just mean a subtle expression of such an emotion, e.g. a gesture, a facial expression, denial of communication, etc. Even if the norm violator would not suffer any material loss, the effectiveness of sanctions is still ensured through the emotion of shame, which a norm violator will suffer when perceiving the sanction "as a vehicle for the emotions of contempt or disgust" [8, p.146]. Also, the mechanisms described by Elster have in part been validated empirically [9, 12, 5, 1].

Damasio provides evidence for the utility of social emotions in view of the internalization of social norms [4]. His somatic-marker hypothesis explains the role of social emotions in categorizing social situations experienced in individual socialization and ontogeny, in which social and personal experiences are to a particularly high extent emotional. Actors experience a wide range of (social) emotions (e.g., sympathy, pride, contempt, shame) which might be induced by punishment or reward. Gradually, experienced situations are categorized according to their intrinsic structure, components, and their significance in terms of prevailing goals and beliefs. Hence, the resulting conceptual categories are connected with the mechanisms responsible for emotion generation. The intrinsic structure of a social situation, options for action and probable future outcomes are more and more associated with corresponding emotions. The subsequent occurrence of a situation of the same category then more the less automatically induces the respective emotional state [4]. "This arrangement allows us to connect categories of social knowledge – whether acquired or refined through individual experience – with the innate, gene-given apparatus of social emotions" [4, p.147]. Moreover, it is demonstrated how cooperation – as an epitome of the norm of reciprocity – can be stimulated and in fact be reinforced by positive emotions [4, p.156].

In view of the interaction of emotion, norms, and normative behavior we can sum up that shame and contempt in particular serve as vehicles for maintaining norms by promoting normative behavior and avoidance of adverse consequences. Compliance with social norms therefore not necessarily arises as a consequence of anticipating a loss of material resources through sanctions. Rather, it is (also) the result of internalized strategies to prevent emotion-based sanctions (e.g., contempt, disdain, detestation or disgust in the punisher) that would entail negative emotions (e.g., shame, guilt, embarrassment) in the violator. Similarly, emotion-based rewards (by way of admiration, appreciation or approval expressed by observers of normative behavior) reliably cause positive emotions (e.g., contentment, satisfaction, pride) in the conforming actor. In consequence, the anticipation of positive emotions facilitates the internalization of social norms. In what follows, we give a sketch of possible formalization of these views and represent them with a Petri Net model of emotion-based norm enforcement and maintenance.

3. PETRI NETS MODELING

In this section we follow a socionic approach [12] to represent the main arguments in a Petri Net based model. Figure 1 shows a Petri Net representation of sanctioning non-complying behavior by means of social emotions, as suggested in [7] and [8]. Rectangles (transitions) represent actions, whereas circles (places) denote available or unavailable resources or conditions that might be met. Arcs directed from transitions to places represent preconditions for action, whereas arcs directed from places to transitions represent an action's outcome. A firing transition (i.e., an action that is implemented) will remove resources or conditions (tokens) from places and insert them into some other place.

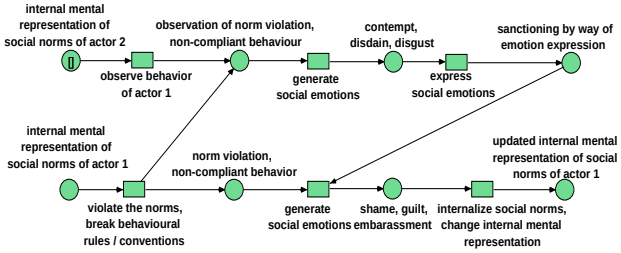

Fig. 1. Internalization [4, 8]

The bottom thread of the net represents the violator's (1) course of action, the upper thread that of an observer / punisher (2). Actor 2 observes the behaviour of actor 1 and perceives a norm transgression (transition "violate the norms, break behavioral rules / conventions"). Social emotions of contempt, disdain or disgust are elicited (transition "generate social emotions") and their expression (transition "express social emotions") constitutes the sanctioning of a norm violator [6, 8] in giving rise (place "sanctioning by way of emotion expression") to negative emotions in the violator (transition "generate social emotions") and induce states of shame, guilt or embarrassment. To integrate the concept of emotion-driven internalization of social norms [4] (s.a.), we extend this model with an additional transition "internalize social norms, change internal mental representation", indicating that certain negative emotions (s.a.) become associated with a situation of norm violation. Occurrence of a similar event would then automatically induce the associated emotions [4]. Associations between negative emotion and norm transgression thus foster the (re-)internalization and reinforcement of social norms and at the same time also update the corresponding internal representation of the norm ("updated internal mental representation of social norms").

Reinforcement of norm-compliant behavior through positive emotions [4, 7, 8] can be modeled in analogy: situations of norm-compliance are associated and internalized along with positive emotions, motivating an actor to seek out situations in which compliance with the (updated) internal representation of a norm leads to intrinsically rewarding positive emotions. Referring to the problem of agent cooperation, this mechanism can be exemplified as cooperative behavior based on adopted social norms within a multi-agent society. In this case, cooperation is mutually rewarded

with different positive emotions, i.e. acknowledgement, gratefulness, or admiration on one side, and pride or satisfaction on the other side. Incentives to cooperate are reinforced on the micro level within a single agent, consequently also reinforcing the norm of cooperation on the macro level of an entire MAS.

Figure 2 integrates processes of sanctioning and reinforcement in a general Petri Net model of emotion-driven norm enforcement and maintenance on the macro level by means of internalization of norms and emotions on the micro level. Here, a powerful variant of Petri Nets – a reference net [17] is used. A notable property of a reference net is the possibility that tokens located on a place in a net (the system net) may again be a reference net (an object net) (or some arbitrary Java object) [17]. Further details on reference nets modeling are omitted here (see [11] for details).

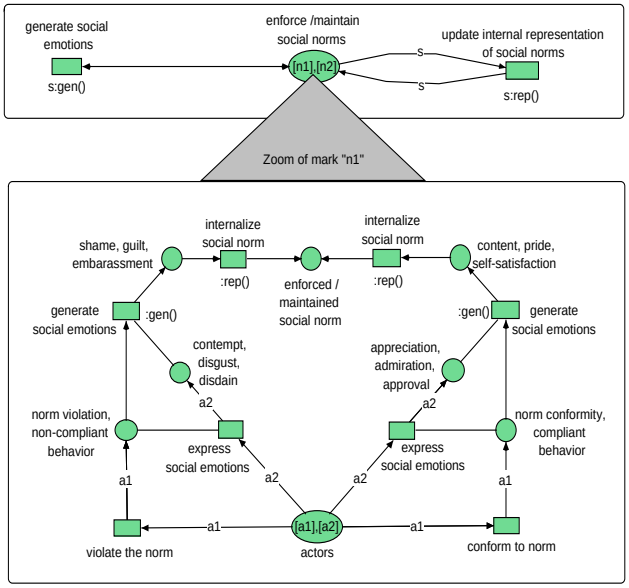

Fig. 2. Emotion-based enforcement and maintenance of norms

4. DISCUSSION

Modeling the social functions of emotion is supposed to improve MAS, e.g. in view of alternative coordination solutions and structuration. Our approach allows modeling the interrelation of social norms and emotion at different levels of abstraction. On the one hand, norms can be modeled on the micro-level, reflecting implications of the internal representation of norms for agents' de facto behavior. On the other hand, norms can also be modeled on the macro-level, thereby focusing on the social mechanisms of norm enforcement and maintenance. A detailed analysis of these approaches and their implications for the dynamics of MAS are clearly beyond the scope of this paper. We basically aimed to show the general role of emotions in enforcing norms in MAS and to evaluate the Petri Net modeling formalisms in view of the interrelation of norms and emotion. The quantitative representation and functions of emotion in agent architectures are expressly *not* addressed herein. Embedding the models into our general multi-agent framework MULAN is the subject of future work.

5. REFERENCES

[1] Berthoz, S., Armony, J.L., Blair, R.J., Dolan R. J. An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain*, 125(8), 2002, 1696-1708.

[2] Castelfranchi, C. The Theory of Social Functions: Challenges for Computational Social Science and Multi-Agent Learning. *Cognitive Systems Research*, 2001, 2(1), 5-38.

[3] Conte, R., Castelfranchi, C. Norms as mental objects. In *Proc. of the 5th Europ. Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW '93)*. Berlin: Springer, 1995, 186-196.

[4] Damasio, A. R. Looking for Spinoza. New York: Harvest, 2003.

[5] Eisenberger, N.I., Lieberman, M.D. Why rejection hurts: a common neural alarm system for physical and social pain. *Trends in Cognitive Sciences*, 8(7), 2004, 294-300.

[6] Elias, N. The Civilizing Process. Oxford: Blackwell, 1978.

[7] Elster, J. Rationality and the Emotions. *The Economic Journal*, 106(438), 1996, 1386-1397.

[8] Elster, J. *Alchemies of the Mind.* Cambridge: Cambridge University Press, 1999.

[9] Fehr, E., Fischbacher, U. Social Norms and Human Cooperation. *Trends in Cognitive Science*, 8(4), 2004, 185-190.

[10] Koehler, M., Moldt, D., Roelke, H. A Discussion of Social Norms with Respect to the Micro-Macro Link. In *Proc. of the 2nd Int. Workshop on Regulated Agent-Based Social Systems (RASTA'03)*, 2003.

[11] von Luede, R., Moldt, D., Valk, R. (Eds.). *Sozionik: Modellierung soziologischer Theorie*. Muenster: Lit, 2003.

[12] de Quervain, D.J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr A. The Neural Basis of Altruistic Punishment. *Science*, 305, 2004, 1254-1258.

[13] von Scheve, C., Moldt, D. Emotion: Theoretical Investigations and Implications for Artificial Social Aggregates. In: G. Lindeman, D. Moldt and P. Paolucci (Eds.) *Regulated Agent-Based Social Systems*. Berlin: Springer, 2004, 189-209.

[14] von Scheve, C., Moldt, D., Fix, J., von Luede, R. My Agents Love to Conform: Norms and Emotion in the Micro-Macro Link. *Computational & Mathematical Organization Theory*. [in press].

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

[16] Staller, A., Petta, P. Introducing Emotions into the Computational Study of Social Norms: A First Evaluation. *Journal of Artificial Societies and Social Simulation*, 4(1), 2001. http://soc.surrey.ac.uk/JASSS/4/1/2.html.

[17] Valk, R. Petri Nets as Token Objects. An Introduction to Elementary Object Nets. In *Proc. of Application and Theory of Petri Nets*. Berlin: Springer, 1998, 1-25.