Introducing Emotions into the Computational Study of Social Norms

Alexander Staller and Paolo Petta
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria (EU)
{alexs, paolo}@ai.univie.ac.at

Abstract

We argue that modelling emotions among agents in artificial societies will further the computational study of social norms. The appraisal theory of emotions is presented as theoretical underpinning of Jon Elster's view that social norms are sustained not only by material sanctions but also by emotions such as shame and contempt. Appraisal theory suggests the following twofold relationship between social norms and emotions: First, social norms play an important role in the generation of emotions; second, emotion regulation depends heavily on the influence of social norms. Based on these insights, we present an emotion-based view on the influential study by Conte and Castelfranchi (1995); without mentioning emotions, they argue that a function of social norms is aggression control. Appraisal theory offers a principled framework for the development of TABASCO, a three-layer agent architecture incorporating social norms. At the macro level, the computational study of social norms can profit by economic and sociobiological theories, which suggest that emotions play an important role in sustaining norms of cooperation and reciprocity. We show how appraisal theory can serve as a link between the macro and micro levels, and summarize the potential benefits from the development of TABASCO.

1 Introduction

Imagine you are invited to dinner. You think this will be an informal event and put on your jeans. However, you soon realize that you are the only guest who is not wearing a dinner jacket or an evening dress. The other guests look contemptuous and avoid talking to you. You feel the tendency to hide, which is a sign of being ashamed.

This example suggests that the violation of a social norm can trigger emotions such as contempt and shame. In this paper, we will elaborate on the relation between social norms and emotions and argue that the computational study of social norms can profit by modelling emotions among agents in artificial societies. In section 2 we will present an emotion-based definition of social norms by Elster (1989). Section 3 is devoted to the appraisal theory of emotions, suggesting that social norms play in important role both in emotion generation and emotion regulation. Appraisal theory provides us with the theoretical underpinning to present an emotion-based view of the study by Conte and Castelfranchi (1995) in section 4. In section 5 we will outline TABASCO, our appraisal-based agent architecture, and present theoretical considerations for incorporating social norms into TABASCO. Section 6 contains a review of economic and sociobiological theories suggesting that emotions play an important role in sustaining norms of cooperation and reciprocity. In section 7 we suggest that appraisal theory can serve as a link between the macro and micro levels. Section 8 concludes the paper by summarizing how the computational study of social norms could benefit from the development of the TABASCO architecture.

2 An Emotion-Based Definition of Social Norms

The example in the introduction suggests that emotions such as contempt and shame play an important role in sustaining social norms. Elster (1996, 1999) has taken this view. He defines social norms as injunctions to behaviour with the following features:

First, social norms are *not outcome-oriented*. In the simplest case they are of the type 'Do X' or 'Do not do X'. If the imperative expressed by a social norm is conditional, then it is not future-oriented. For example it is of the type 'If others do Y, then do X'. By contrast, rational action is concerned with outcomes. A rational, self-interested actor follows the maxim 'If you want to achieve Y, do X'.

Second, for norms to be *social*, they must be shared by other people. Some norms are shared by all members of the society, while other norms are more group-specific. Another respect in which norms are social is that other people are important for enforcing them through sanctions.

Third, social norms are not only sustained by the sanctions of others, but also by *emotions*. The violation of a social norm can trigger negative emotions such as shame or guilt in the norm violator, even if nobody can observe the norm violation. So emotions arise as negative internal consequences of a norm violation and thus sustain social norms in addition to external sanctions.

On this account, emotions do not seem to be a necessary part of a system of social norms. The enforcement of social norms appears to be overdetermined by sanctions and emotions. But Elster (1996, 1999) argues that emotions are crucial for the operation of sanctions. A person who is imposing sanctions on the norm violator is driven by emotions such as contempt or disgust. A sanction may be just a subtle expression of such an emotion, e.g. a facial expression. Even if the norm violator does not suffer any material loss, the sanction is still effective because the norm violator "will see the sanction as a vehicle for the emotions of contempt or disgust and suffer shame as a result" (Elster , 1999, p. 146). The introductory example is a case in point.

Elster's view presupposes that social norms play an important role in the *generation* of emotions such as contempt and shame. In addition, he notes that emotions and their expression may be *regulated* by social norms. As an example he puts forward the norm against laughing at funerals (Elster , 1996).

Is there any theoretical support for this twofold relation between social norms and emotions? Indeed, appraisal theory – especially Frijda's (1986) approach – explicitly deals with the role of social norms in the generation and regulation of emotions. In the next section, we describe appraisal theory in more detail.

3 The Appraisal Theory of Emotions

After having long been dismissed as irrational and of no utility, emotions are now seen as a key element in successful coping with a non-deterministic, dynamic, and social environment. Appraisal theory emphasizes that this coping depends on the continuous monitoring of the relationship between the individual and the environment. Its central tenet "is the claim that emotions are elicited and differentiated on the basis of a person's subjective evaluation or appraisal of the personal significance of a situation, object, or event on a number of dimensions or criteria" (Scherer , 1999, p. 637). Thus, appraisal theory explains why the same event can give rise to different emotions in different individuals, or even in one and the same individual at different times. Conversely, appraisal theory offers a framework for the identification of the conditions for the elicitation of different emotions, as well as for understanding what differentiates emotions from each other.

3.1 Appraisal Criteria

Many theorists have been trying to specify the criteria according to which a situation is appraised (Roseman, 1984; Scherer, 1984; Smith and Ellsworth, 1985; Frijda, 1986; Ortony et al., 1988; Lazarus, 1991). There is a high degree of consensus with respect to these criteria. According to van Reekum and Scherer (1997, pp. 259-260), these include "the perception of a change in the environment that captures the subject's attention (novelty and expectancy), the perceived pleasantness or unpleasantness of the stimulus or event (valence), the importance of the stimulus or event to one's goals or concerns (relevance and goal conduciveness or motive consistency), the notion of who or what caused the event (agency or responsibility), the estimated ability to deal with the event and its consequences (perceived control, power or coping potential), and the evaluation of one's own actions in relation to moral standards or social norms (legitimacy), and one's self-ideal."

The postulate of appraisal theory is that specific profiles of appraisal outcomes on these criteria determine the nature of the ensuing emotion. Scherer (1999, p. 639) provides a table of theoretically contended appraisal profiles for anger/rage, fear/panic, and sadness.

3.2 The Appraisal Process

The description of the appraisal criteria in abstract, conceptual terms, often represented as a series of questions to be evaluated, led many critics to assume that the appraisal process is necessarily deliberate and conscious. For example, Zajonc (1980) criticized the "exaggerated cognitivism" of appraisal theory. In response to this criticism appraisal theorists pointed out that the appraisal process largely occurs nonconsciously and involves perceptual processing. The fear of a tiger jumping out of the bush is certainly not elicited by a conscious evaluation of appraisal criteria, but by fast perceptual processes. The appraisal process involves perceiving the "affordance" (Gibson, 1979) of stimulus events for one's coping activities (Smith and Lazarus , 1990; Frijda , 1993).

An affordance is defined by Gibson (1979, p. 127) as "what it offers the animal, what it provides or furnishes, either for good or ill." The general idea is that an animal actively perceives meaning in the environment without further interpretative cognitive processing. So there is a direct coupling between perception and action. McArthur and Baron (1983) apply the affordance concept to social perception, e.g. to emotion perception, impression formation, and causal attribution.

Leventhal and Scherer (1987) include perceptual processing in their model of the appraisal process. They suggest a hierarchical processing system consisting of three levels: sensory-motor, schematic, and conceptual. The sensory-motor level is based on innate hard-wired feature detectors which can give rise to emotional reaction directly. The schematic level is based on schema matching. The conceptual level involves reasoning and inference processes that are abstract, active, and reflective.

Building on the model by Leventhal and Scherer (1987), Smith and Kirby (2000) suggest a model of the appraisal process in which perceptual processing is complemented by associative processing (i.e., schematic processing) and reasoning (i.e., conceptual processing). Associative processing is a fast, automatic, parallel, and memory-based mode of processing. As memories of previous experiences are activated, appraisal meanings associated with them are activated automatically. In contrast, reasoning is a relatively slow, controlled, and serial process that actively constructs appraisal outcomes. A novel feature of this model is the existence of so-called *appraisal detectors*. They monitor appraisal information generated through associative processing and reasoning, in addition to perceptual information, and generate an emotional reaction. The appraisal detectors are assumed to model the function of the amygdala, which plays an important role in the elicitation of fear (LeDoux , 1996) and presumably of other emotions as well.

The view of appraisal as a multi-level process corresponds to the recent trend towards multi-level theories of cognition-emotion relations in the areas of clinical psychology, neuropsychology, and the study of memory (Teasdale , 1999). Van Reekum and Scherer (1997) discuss the pertinence of such theories for a model of the appraisal process in more detail.

3.3 The Emotion Process

Throughout the rest of section 3 we follow Frijda (1986), a main proponent of appraisal theory. All citations refer to Frijda (1986).

Appraisal is the first step of the emotion process. For successful coping with the environment the appraisal outcome must have an effect on the actions of the individual. But appraisal does not lead to action directly. Instead, appraisal is followed by an impulse, i.e., the instigation of an action tendency. Action tendencies “are states of readiness to achieve or maintain a given kind of relationship with the environment. They can be conceived of as plans or programs to achieve such ends, which are put in a state of readiness” (p. 75). For example, the action tendency of fear is avoidance. An impulse involves shifts in control over behaviour, attention, and resources that are referred to as the “control precedence” feature of emotion. Frijda et al. (1989) have established significant relations between particular appraisal patterns and action tendencies. Thus emotions can be regarded both as experiences of forms of appraisal and as states of action readiness. The final step of the emotion process is the generation of cognitive or overt action, possibly in the form of mostly expressive behaviour such as facial expressions.

Frijda emphasizes the importance of emotion regulation. All steps of the emotion process sketched so far are subject to regulatory processes. Regulatory processes include: the modification of appraisal, e.g. by reappraising a situation; impulse control, e.g. the suppression of an action tendency; and the modification of action, e.g. by attenuating or replacing expressive behaviour.

Important for the instigation of regulatory processes are signals of aversive outcomes of unrestrained emotional behaviour. These outcomes can be external or internal. An example of an aversive external outcome is punishment, “when the environment retaliates, envies, disapproves, or despises because of emotions shown” (p. 409). Signals of aversive internal outcomes are “the calls of conscience and the sense of propriety” (p. 409).

In sum, emotional response is under dual control. Generative processes are modulated by regulatory processes. The next two sections highlight the importance of social norms for both emotion generation and emotion regulation.

3.4 The Role of Social Norms in Emotion Generation

Social norms enter the process of emotion generation during appraisal. The definition of the appraisal criterion “legitimacy” (see section 3.1) is based on social norms. Many emotions are contingent upon adherence or violation of social norms. Examples are “comfort in one’s sense of propriety, pride in one’s outstanding achievements, admiration for those of others; shame and guilt upon one’s own infringements and distrust, anger, and indignation upon those of others” (p. 311). This list makes clear that to differentiate these emotions, the appraisal criterion “agency or responsibility” is necessary. Shame and guilt are contingent upon a norm violation by oneself, while contempt and anger are contingent upon a norm violation by another.

Scherer (1988, p. 112) provides a table of the complete appraisal patterns for some major emotions including shame, guilt, anger, contempt, and pride.

3.5 The Role of Social Norms in Emotion Regulation

Social norms are crucial for the instigation of emotion regulation. As mentioned in section 3.3, signals of aversive external or internal outcomes of unrestrained emotional behaviour instigate regulatory processes. Punishment was given as an example of an aversive external outcome. Of course, the violation of social norms is a major reason for punishment through sanctions.

Social norms also underly “the calls of conscience and the sense of propriety” signaling aversive internal outcomes. These signals consist in the anticipation of emotions such as shame or guilt that would be elicited by a norm violation.

Very important for the instigation of emotion regulation are social norms regarding the appropriateness of emotions and their expression. Hochschild (1983) focuses on culture-specific “feeling rules” and “expression rules.” She shows that a good deal of our emotional life consists of “emotion work” that brings our emotions and their expression in line with these normative prescriptions. The rule against laughing at funerals mentioned in section 2 is an example of such a prescription.

Ekman and Friesen (1975) extensively discuss culturally defined “display rules” prescribing appropriate expressive behaviour. They distinguish four strategies for putting display rules into practice: “minimization,” i.e., miniaturizing the expression; “maximization,” i.e., exaggerating the expression; “masking,” i.e., adopting a neutral expression; and “substitution,” i.e., expressing a different emotion.

A considerable part of emotion socialization in childhood is devoted to the acquisition of norms regarding the appropriateness of emotions and their expression. Saarni (1993) distinguishes five methods of emotion socialization: direct instruction, contingency learning, imitation, identification with role models, and communication of expectancies.

4 An Emotion-Based View on an Influential Study

Conte and Castelfranchi (1995) realized that previous work in Artificial Intelligence (Shoham and Tennenholtz, 1992a,b) had a very restricted view of norms. Based on game theory, norms were seen essentially as conventions permitting or improving coordination among agents. Conte and Castelfranchi (1995) conducted a study to investigate another function of norms: the control of aggression among a population of agents. This research has been very influential, forming the basis of several studies (Walker and Wooldridge, 1995; Castelfranchi et al., 1998; Saam and Harrer, 1999). In the following, it is described briefly:

Agents perform some elementary routines for surviving in a situation of food scarcity (e.g., moving, eating, attacking an eating agent). Each agent has a strength, which is increased by eating and decreased by moving and attacking. In one condition, each agent owns a number of food items. All agents follow a normative strategy for aggression control: They do not attack agents eating their own food, i.e., they comply with the “finder-keeper” norm. In another condition, all agents follow a utilitarian strategy for aggression control: They do not attack eating agents whose strength is higher than their own. The normative strategy has been found to reduce aggression (i.e., the number of attacks) to a much greater extent than the utilitarian strategy.

Conte and Castelfranchi (1995) studied the function of the “finder-keeper” norm as a macro-social object. So the agents were deliberately kept as simple as possible and could just execute elementary routines. The term “aggression” simply denotes the execution of the “attack” routine.

How could agents be implemented that more accurately model the psychological processes underlying aggression control in humans? To this end, we point out that aggressive behaviour is a main example of emotional behaviour. Neither Conte and Castelfranchi (1995) nor the authors of the follow-up studies ever mention emotions.

Appraisal theory offers a detailed account of the processes underlying the generation and control of aggressive behaviour in humans:

Frijda calls the action tendency underlying aggressive behaviour “agonistic” (Frijda , 1986, p. 88). The agonistic action tendency covers attack and threat. The emotion corresponding to this action tendency is anger. The agonistic action tendency is generated by the appraisal that the satisfaction of a concern is obstructed. The end state of the agonistic action tendency is the removal of this obstruction.

A basic concern of a living being is the optimal state of feeding. Another person in possession of scarce food is appraised as obstructing the satisfaction of this concern. This appraisal leads to the generation of the agonistic action tendency. If this action tendency is not suppressed, overt aggressive behaviour (e.g. an attack) is generated.

Aggression control can thus be viewed as an example of impulse control, namely the suppression of the agonistic action tendency. In section 3.3 we mentioned that regulatory processes can be instigated by signals of aversive external or internal outcomes of unrestrained emotional behaviour. Punishment was given as an example of an external aversive outcome. When the person in possession of food is stronger than oneself, retaliation can lead to punishment for unrestrained aggression. If the “finder-keeper” norm is in force, aggression control is either due to the anticipation of punishment through sanctions or due to “the calls of conscience and the sense of propriety,” i.e., the anticipation of shame or guilt as aversive internal outcomes (see section 3.5).

This short account of the generation and control of aggression suggests that appraisal theory can guide the development of a psychologically more plausible agent architecture. In the next section we will sketch our attempts to flesh out TABASCO[1], our appraisal-based agent architecture.

5 The Development of TABASCO

5.1 Existing Appraisal-Based Architectures

The majority of the current appraisal-based architectures used to engender emotional competence in software agents include some reified representation of a finite number of discrete emotional states through which all emotional processing is explicitly routed. Well-known examples of such architectures are the Affective Reasoner (Elliott, 1992) and Em (Reilly, 1996), the emotional component of the Tok architecture developed in the Oz project (Bates et al., 1992). Both architectures rely on the theory by Ortony et al. (1988), which quickly has become

[1] The name stands for “A **T**ractable **A**ppraisal-**B**ased **A**rchitecture for **S**ituated **Co**gnizers.”

the most popular "reference model" of appraisal used in agent architectures.

The reification of emotional states results in the implementation of a full explicit mapping from these states to entailed effects, including internal processing and externally observable overt behaviour. The characteristics of systems engineered according to such a shallow approach are well known from the traditional research area of expert systems in artificial intelligence: the merits of rather straightforward design–for moderate ruleset size–and precisely known coverage stand against a number of problems, including brittleness of system behaviour surfacing with every occurrence of any situation not explicitly anticipated at implementation time; laboriousness of system extension; and the issue of overall system consistency: as the agent's behaviour is governed by a large collection of independent rules–as opposed to a small set of generating principles–it falls into the responsibility of designers and implementors to ensure that with a growing body of incorporated knowledge the system remains free of inconsistencies and continues to perform in a desired and coherent fashion.

Besides the problems of brittleness and consistency, reification of emotions as identifiable system components and routing of all processing through these entities engenders the problem of how to proceed from these emotions for further system processing, leading to the adoption of ad-hoc constructions of dubious validity.

The Affective Reasoner is an architecture for agents in a multi-agent world with the capability of abstract, domain-independent reasoning about emotion episodes. Such an architecture runs into serious problems when deployed in interactive virtual scenarios: to be effective in such applications, affective reasoning has to have appropriate access to pertinent information about and from the world, and has to be able to influence the overt external behaviour as well as the internal information processing of an agent. The only means to achieve this is to fully integrate emotional competence into an architecture which in turn has to be adapted to the environment in which the agent is situated.

5.2 The TABASCO Architecture

TABASCO is an attempt to overcome the problems stated above. It has first been adumbrated by Staller and Petta (1998). TABASCO integrates the three-level model of the appraisal process (see section 3.2) into a three-layer architecture for software agents situated in a virtual environment. Three-layer architectures have emerged as robust, widely adopted solutions to fundamental aspects of the realization of situated agents (Gat , 1997). Emotions are not modelled as reified entities, but as an adaptive process related to the agent-environment interaction, with the appraisal process and the execution of action tendencies as main components. Action tendencies provide a principled way of classifying the behavioural repertory of an agent in classes that share specific expressive characteristics, obviating the need of ad-hoc solutions. The implementation of the emotion process does not follow exactly Frijda (1986) who proposes a sequential process (see section 3.3). Our layered architecture has the advantage that the agent can respond reactively to events in the environment without having to execute a sequential process with action generation as the last step.

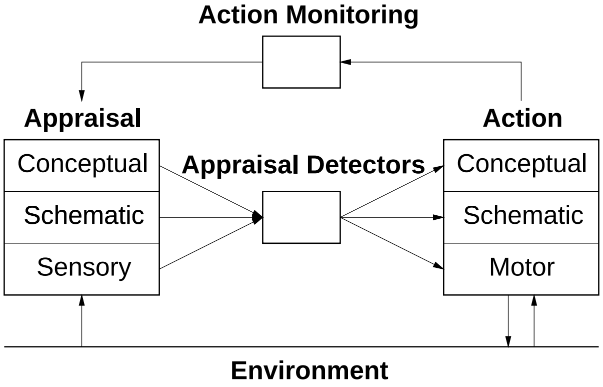

Figure 1: The TABASCO Architecture

In the following, we sketch the conceptional design, which is shown in figure 1.

The psychological idea underlying TABASCO is that the distinction between sensory-motor, schematic, and conceptual processing does not only apply to appraisal, but also to the generation of action, as proposed originally by Leventhal in his "perceptual-motor theory of emotion" (Leventhal, 1984).

The **Appraisal** component processes environmental stimuli and models the appraisal process based on the three-level theory of the appraisal process (see section 3.2).

The **Action** component models long-term planning processes at the conceptual level, the generation of action tendencies at the schematic level, and action generation at the motor level.

The **Appraisal Detectors**, suggested by Smith and Kirby (2000), detect and combine the appraisal outcomes, and instigate processes in the **Action** component: planning processes, action tendencies, and actions.

The **Action Monitoring** component monitors the planning and execution processes in the **Action** component and sends the results to the **Appraisal** component, where it is integrated into the appraisal process.

So far we have mainly been concerned with designing an architecture for emotion generation. The whole range of regulatory processes described by Frijda (1986) has not yet been incorporated into the TABASCO architecture. But e.g. impulse control can be modelled by letting the planning processes at the conceptual level intervene in the processes at the schematic level so that action tendencies can be prevented from being executed. It is also possible that results of **Action Monitoring** that are sent to the **Appraisal** component lead to a reappraisal of a situation. For example, the execution of actions without success may lead to a reappraisal of the appraisal criterion "perceived control, power or coping potential" (see section 3.1).

A version of TABASCO has been implemented for the control of a synthetic character interacting with users in an immersive interactive virtual environment (Petta, 1999; Petta et al., 1999). The implementation is based on 3T (Bonasso et al., 1999), a three-layer agent architecture with the layers *deliberation*, *sequencing*, and *reactive skills*. The deliberation layer consists of a planner and corresponds to conceptual processing. The sequencing layer corresponds to schematic processing. It is based on the Reactive Action Packages (RAPs) system. (Firby, 1989). The reactive skills correspond to sensory-motor processing. So far we have concentrated on implementing the generation and management of action tendencies based on RAPs.

Another line of research along which we are trying to flesh out TABASCO is the implementation of FORREST (Petta et al. , 2000), an agent situated in multi-user realtime text-based environments known as MUDs (Curtis , 1992). FORREST is an expansion of the Colin MUDbot (Mauldin , 1994). The C code of Colin was complemented with a fairly accurate implementation of Frijda's sequential model of the emotion process (see section 3.3). Most of the emotion process takes place in a module written in NASA's expert system programming shell, CLIPS (1993). The rule-based implementation of a sequential emotion process forms the basis of the conceptual level of TABASCO. The next step towards a realization of TABASCO is the implementation of associative processing at the schematic level. With respect to social simulations, a MUD has the advantages that it is already designed as a multi-user system, in which an arbitrary number of agents and users can interact with each other and share equal "symbolic" access to the environment. We plan to exploit these facts in future social simulations.

Social norms have not yet been incorporated in our implementation. In the next section, we present some first theoretical considerations for the incorporation of social norms into TABASCO.

5.3 Incorporating Social Norms into TABASCO

5.3.1 Emotion Generation

In section 3.4 we pointed out that social norms enter the process of emotion generation during appraisal. For the evaluation of the "legitimacy" criterion it must be determined whether a social norm has been violated.

The implementation of the normative reasoning processes underlying the "legitimacy" check at the conceptual level of the **Appraisal** component can certainly profit by research on deontic logic. For example, Conte et al. (1999) present a logical framework for the specification of "norm-autonomous" agents. Their approach is based on explicit representations for goals, intentions, and beliefs. A norm is conceptualized as an obligation on a given set of agents to accomplish or abstain from a given action. Incorporating norms into agent architectures based on logic is a common approach. However, it would be problematic to base an architecture for a situated agent solely on a logical framework. Wooldridge and Jennings (1995) point out several problems of such a "deliberative" agent architecture, e.g., the problem of maintaining an explicitly represented, symbolic world model in a rapidly changing environment. In contrast, a layered architecture such as TABASCO allows the agent to react directly to changes in the environment without relying on a world model.

In TABASCO the evaluation whether a norm has been violated is not restricted to the conceptual level. The schematic level is also involved. It has even been hypothesized by Leventhal and Scherer (1987) and proponents of other multi-level theories (Teasdale , 1999) that the associative processes at the schematic level are necessary for emotion elicitation, while the "cold" cognitions at the conceptual level have only a subsidiary function. In the following, we present some theoretical ideas on the incorporation of social norms into the schematic level of the **Appraisal** component:

Leventhal and Scherer (1987) use social schemas for the conceptualization of social norms at the schematic level, but do not provide any detail. Social schemas are a central concept of social cognition (Fiske and Taylor , 1991; Augoustinos and Walker , 1995).

Important social schemas are event schemas (scripts), which specify the appropriate behavioural sequence of events, e.g. of eating in a restaurant. Scripts were introduced by Schank and Abelson (1977) to account for the human ability to understand more than was being referred to explicitly in a sentence by explaining the organization of implicit knowledge of the world one inhabits. When we hear the sentence "John ordered sushi but he didnt like it," the restaurant script allows us to infer that this sentence is about eating.

Schank (1982) modified his view of scripts. The starting point of his theory is the conceptualization of a script as a dynamic memory structure. A script is not an unchangeable data structure, but changes over time by storing the memories of episodes deviating from the script. For example, a person who has never been in a Japanese restaurant uses the restaurant script to form expectations about what will happen. Receiving chopsticks instead of a fork is an expectation failure. This expectation failure is stored at the script juncture where it occurred, so that the next time the person receives chopsticks the memory is retrieved and made available for use. Schank calls this conception of memory failure-driven memory.

This conception of scripts is useful for implementing social norms and the detection of norm violations at the schematic level of the **Appraisal** component. The script defines the sequence of actions prescribed by the norm, while the episodes deviating from the script correspond to norm violating episodes. So the detection of a norm violation simply amounts to reminding of expectation failures.

In fact, the implementation of the schematic level of the **Appraisal** component can generally be based on the conception of scripts as organizing memories of expectation failures. Unexpected events are exactly the kind of events that can give rise to emotions. Based on the model by Smith and Kirby (2000) (see section 3.2), the memories of these events are directly associated with the respective appraisal patterns. Then appraisal at the schematic level merely involves reminding deviations from the script and following the link to the associated appraisal pattern.

Schank (1982) further elaborates his theory based on so-called memory organization packets (MOPs). MOPs cover more general knowledge than scripts. For example, Schank proposes the existence of a MOP for a professional office visit that applies to visits to a doctor and to a lawyer equally, while these events are covered by separate scripts. This theory has formed the basis of case-based reasoning (Kolodner , 1993) and can account for more results of memory research than scripts alone. The final implementation of the schematic level of the **Appraisal** component may be based on this theory and case-based reasoning techniques, e.g. for case representation and indexing.

5.3.2 Emotion Regulation

In section 3.5 we pointed out that social norms play an important role in emotion regulation because a norm violation through unrestrained emotional behaviour can be the reason for punishment (an aversive external outcome) or for the generation of emotions such as shame and guilt (an aversive internal outcome). Crucial for the instigation of regulatory processes is the ability to anticipate these aversive outcomes. This ability largely relies on learning. For example, if a certain emotional behaviour has led to negative consequences, a memory of this experience can be stored in memory and used for the timely instigation of regulatory processes in similar situations in the future.

How can such a memory-based instigation of regulatory processes be modelled in TABASCO? As an example, we focus on impulse control based on the memory that a previous execution of an action tendency led to guilt. We cannot specify the computational processes exactly, but we outline which components of TABASCO may be involved in an implementation:

In section 5.2 we suggested that impulse control can be modelled by letting processes at the conceptual level of the **Action** component prevent action tendencies generated at the schematic level from being executed. In order to suppress an action tendency, the conceptual level of the **Action** component must have access to the memory of a similar situation in which the action tendency was executed. This memory is located at the schematic level of the **Appraisal** component and may be represented and retrieved based on Schank's (1982) memory theory or case-based techniques. The association of this situation with guilt is represented by means of an associative link between the memory of the situation and the appraisal pattern of guilt. Currently, the design as shown in figure 1 does not contain a direct connection between the schematic level of the **Appraisal** component and the conceptual level of the **Action** component, but there is no reason against it.

Our emotion-based view of the study by Conte and Castelfranchi (1995) presented in section 4 suggests that appraisal theory could guide the development of agents that model the processes of aggression control in a psychologically more plausible way. In TABASCO the behaviour of agents complying with the “finder -keeper” norm could be modelled by the processes of memory-based impulse control outlined above. The action tendency to be suppressed is the agonistic action tendency.

Our conception of a memory-based instigation of regulatory processes in TABASCO is a way of modelling what Frijda refers to as “the calls of conscience and the sense of propriety” (see section 3.3). Even if no external punishment is expected, regulatory processes are instigated based on memories of situations associated with appraisal patterns of guilt or shame.

Emotion regulation based on memories associated with outcomes of unrestrained emotional behaviour is an instance of contingency learning, which has been identified by Saarni (1993) as a mechanism of emotion socialization (see section 3.5). The other mechanisms such as direct instruction, imitation, identification with role models, and communication of expectancies certainly require more complex cognitive processes, and proposing how to model them is beyond the scope of this paper.

6 Functions of Emotions at the Macro Level

Emotions have an important adaptive function for the individual. According to appraisal theory, they support the individual in the satisfaction of concerns or goals by instigating action tendencies. These action tendencies are directed towards establishing or maintaining a certain relationship with the environment. However, the environment is a social environment. Appraisal theory focuses on the internal psychological processes underlying emotions, but remains largely silent about potential social functions of emotions.

In this section we briefly review three theories suggesting that emotions have the important function of sustaining norms of cooperation and reciprocity. These theories do not explicitly claim that emotions sustain social norms, but they share the view that certain emotions (e.g., anger) bring a person to punish a cheater (i.e., a person who failed to cooperate or reciprocate). Under the assumption that norms of cooperation and reciprocity are in force, this amounts to imposing a sanction on a norm violator. Regarding the existence of such norms, it has been hypothesized that the norm of reciprocity is universal, i.e., that it exists in all human cultures (Gouldner , 1960).

6.1 Reciprocal Altruism

Altruistic behaviour benefits another person, while being apparently detrimental to the person performing the behaviour. Helping and sharing food are examples of altruistic behaviour. Trivers (1971) explains altruistic behaviour toward nonkin with a theory of "reciprocal altruism." Based on this theory, people perform an altruistic act in the expectation that the beneficiary will reciprocate in the future.

Trivers argues that emotions play a crucial role in the evolution of reciprocal altruism. For example, "moralistic aggression" has been selected for in order to punish unreciprocating individuals ("cheaters") e.g. by cutting off all future altruistic acts. Guilt has been selected for in order to motivate the cheater to make up for his misdeed and thus to continue reciprocal relationships. Trivers enumerates a number of other emotions that he regards as important for the regulation of the altruistic system.

6.2 Emotions as "Commitment Operators"

Based on Trivers's work, Aubé (1998) proposes that emotions might have emerged to control and manage commitments among members of a society. Aubé borrows the notion of commitment from symbolic interactionism in sociology (e.g., Becker, 1960) and distributed artificial intelligence (e.g., Fikes, 1982). Commitments bind agents together into cooperative behaviour. Aubé calls emotions "commitment operators" that "operate so as to establish or create new commitments (joy, gratitude), protect, sustain, or reinvest old ones (joy, hope, gratitude, pride), prevent the breaking of commitments by self or others (pride, guilt, gratitude, anger), or call on 'committed' others in cases of necessity, danger, and helplessness (sadness, fear)." (Aubé, 1998, p. 15).

Commitments are conceived of as "second-order resources" in addition to vital "first-order resources" such as food. Based on this classification of resources, Aubé suggests a two-layer control system: Needs such as hunger control first-order resources, while emotions control second-order resources.

6.3 Emotions as "Commitment Devices"

Frank (1988) also uses the term "commitment," but his conception of this term differs from Aubé's. Frank proposes that in social dilemmas such as the prisoner's dilemma some emotions, the so-called "moral sentiments," commit a person to act contrary to self-interest. For example, the predisposition to feel guilt commits a person to cooperate, even if cheating were in his material interest. A person with the predisposition to get outraged after having been cheated is committed to punish the cheater, even if it is costly in material terms. So emotions such as guilt and anger act as "commitment devices" that change the material incentives.

But there must be a material gain from having these emotions, otherwise they would not have evolved. Frank proposes that emotional predispositions have long-term material advantages: An honest person with the predisposition to feel guilt will be sought as a partner in future interactions. The predisposition to get outraged will deter others from cheating.

However, others must be able to discern the presence of these emotional predispositions. Frank suggests two ways how this might occur: The first is reputation. The knowledge about the honesty or the vengefulness of a person can be spread among the population. The second way of discerning emotional predispositions is through physical and behavioural clues, such as facial expressions, voice, and posture. Frank discusses the reliability of these clues and the problem of deception, but this discussion is beyond the scope of this paper.

7 Connecting the Macro and Micro Levels

The theories reviewed above focus on the functions of emotions at the macro level, while appraisal theory specifies the processes occurring at the micro level. Is there any connection between these two levels of analysis? Indeed, the macro-level theories make assumptions about micro-level processes that are fully in accordance with appraisal theory.

For example, the theories assume that the experience of having been cheated leads to anger. But what is the psychological process underlying the realization that one has been cheated? It can be thought of as an appraisal process: Having been cheated is appraised as a situation in which the satisfaction of a concern or goal has been obstructed and another agent is responsible for this obstruction. These are the crucial appraisal outcomes for the generation of anger.

The theories also assume that emotions have an influence on actions. For example, Trivers claims that guilt motivates the cheater to make up for his misdeed. This is exactly the action tendency of guilt. Punishing a cheater can be interpreted as due to the agonistic action tendency of anger.

These examples suggest that appraisal theory can serve as a link between the macro and micro levels. Macro-level functions of emotions such as sustaining cooperation and reciprocity depend on the micro-level processes of appraisal and the generation of action tendencies. This insight paves the way for testing the macro-level theories in social simulations with agents that are able to perform appraisal and the generation of action tendencies. TABASCO is a proposal for the implementation of such agents.

8 Conclusion

In this paper, we have tried to show that the computational study of social norms can profit by modelling emotions among agents in artificial societies. We have suggested appraisal theory as the theoretical foundation for endowing agents with emotions. Our TABASCO architecture is a proposal for the development of appraisal-based agents. The computational study of social norms can benefit from the development of TABASCO in the following ways:

- Social norms must be represented in TABASCO. Appraisal theory, especially the three-level theory of the appraisal process can guide the exploration of representations that are not based on logic. We have suggested social schemas, especially scripts, as the basis for this exploration.
- The insight that appraisal and action tendencies can serve as a link between the macro and micro levels paves the way for testing the macro-level emotion theories – which suggest that emotions serve to sustain norms of cooperation and reciprocity – in social simulations with TABASCO agents.
- The account of appraisal theory for emotion regulation through social norms sheds new light on existing research. From the point of view that aggression control is an instance of impulse control, a large part of computational research on social norms has investigated a special case of emotion regulation through social norms. The implementation of regulatory processes in TABASCO leads to a psychologically plausible model of emotion regulation through social norms.

In general, emotions are of paramount importance for the social life of humans and should therefore not be neglected in the study of artificial societies.

Acknowledgements

We would like to thank Robert Trappl for helpful comments on an earlier draft of this paper. This research has been supported by the Austrian Federal Ministry of Science and Transport.

References

M. Aubé. A commitment theory of emotions. In *Emotional and Intelligent: The Tangled Knot of Cognition, Proceedings of the 1998 AAAI Fall Symposium*, AAAI Technical Report FS-98-03, Orlando, FL, 13–18, 1998.

M. Augoustinos and I. Walker. *Social Cognition: An Integrated Introduction*. Sage, London, 1995.

J. Bates, A.B. Loyall, and W.S. Reilly. An architecture for action, emotion, and social behavior. In *Proceedings of the Fourth European Workshop on Modeling Autonomous Agents in a Multi-Agent World*. San Martino al Cimino, Italy, 1992.

H.S. Becker. Notes on the concept of commitment. *American Journal of Sociology*, 66:32–40, 1960.

R.P. Bonasso, R.J. Firby, E. Gat, D. Kortenkamp, D.P. Miller, and M. Slack. Experiences with an architecture for intelligent, reactive agents. In H. Hexmoor (ed.), Special Issue: Software Architectures for Hardware Agents, *Journal of Theoretical and Experimental Artificial Intelligence*, 9(2/3):237–256, 1997.

C. Castelfranchi, R. Conte, and M. Paolucci. Normative reputation and the costs of compliance. *Journal of Artificial Societies and Social Simulation*, 1(3), <http://www.soc.surrey.ac.uk/JASSS/1/3/3.html>, 1998.

R. Conte and C. Castelfranchi. Understanding the functions of norms in social groups through simulation. In G.N. Gilbert and R. Conte (eds.), *Artificial Societies: The Computer Simulation of Social Life*. UCL Press, London, 252–267, 1995.

R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In J.P. Müller, M.P. Singh, A.S. Rao (eds.), *Intelligent Agents V - Proceedings of the Fifth International Workshop on Agent Theories, Architectures, and Languages (ATAL- 98)*. Springer-Verlag, Heidelberg, 99–112, 1999.

P. Curtis. Mudding: Social phenomena in text-based virtual realities. In *Proceedings of the 1992 Conference on Directions and Implications of Advanced Computing*, Berkeley, CA, 1992.

P. Ekman and W.V. Friesen. *Unmasking the Face*. Prentice Hall, Englewood Cliffs, NJ, 1975.

C.D. Elliott. *The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System*. PhD Thesis, Northwestern University, Evanston, IL, 1992.

J. Elster. *The Cement of Society: A Study of Social Order*. Cambridge University Press, Cambridge, UK, 1989.

J. Elster. Rationality and the emotions. *The Economic Journal*, 106(438):1386–1397, 1996.

J. Elster. *Alchemies of the Mind: Rationality and the Emotions*. Cambridge University Press, Cambridge, UK, 1999.

R.E. Fikes. A commitment-based framework for describing informal cooperative work. *Cognitive Science*, 6(4):331–348, 1982.

R.J. Firby. *Adaptive Execution in Complex Dynamic Worlds*. PhD Thesis, Yale University, New Haven, CT, 1989.

S.T. Fiske and S.E. Taylor. *Social Cognition* (2nd edition). McGraw-Hill, New York, 1991.

R.H. Frank. *Passions within Reason: The Strategic Role of the Emotions*. Norton, New York, 1988.

N.H. Frijda. *The Emotions*. Cambridge University Press, Cambridge, UK, 1986.

N.H. Frijda. The place of appraisal in emotion. *Cognition and Emotion*, 7(3/4):357–388, 1993.

N.H. Frijda, P. Kuipers, and E. ter Schure. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, 57(2):212-228, 1989.

E. Gat. On three-layer architectures. In D. Kortenkamp, R.P. Bonasso, and R. Murphy (eds.), *Artificial Intelligence and Mobile Robots*, MIT/AAAI Press, 1997.

J.J. Gibson. *The Ecological Approach to Visual Perception*. Houghton-Mifflin, Boston, 1979.

A.W. Gouldner. The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25:161–178, 1960.

A.R. Hochschild. *The Managed Heart: Commercialization of Human Feeling*. University of California Press, Berkeley, 1983.

J. Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, San Mateo, 1993.

R.S. Lazarus. *Emotion and Adaptation*. Oxford University Press, New York, 1991.

J.E. LeDoux. *The Emotional Brain*. Simon & Schuster, New York, 1996.

H. Leventhal. A perceptual-motor theory of emotion. *Advances in Experimental Social Psychology*, 17:117–182, 1984.

H. Leventhal and K.R. Scherer. The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition and Emotion*, 1(1):3–28, 1987.

M.L. Mauldin. CHATTERBOTS, TINYMUDS, and the Turing Test: Entering the Loebner Prize Competition. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*. AAAI Press/MIT Press, Cambridge, MA, 16–21, 1994.

L.Z. McArthur and R.M. Baron. Toward an ecological theory of social perception. *Psychological Review*, 90(3):215–238, 1983.

A. Ortony, G.L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, UK, 1988.

P. Petta. Principled generation of expressive behavior in an interactive exhibit. In J.D. Velasquez (ed.), *Workshop: "Emotion-Based Agent Architectures" (EBAA'99)*, Third International Conference on Autonomous Agents (Agents '99), Seattle, WA, 94–98, 1999.

P. Petta, A. Staller, R. Trappl, S. Mantler, Z. Szalavari, T. Psik, and M. Gervautz. Towards engaging full-body interaction. In H.-J. Bullinger and P.H. Vossen (eds.), *Adjunct Conference Proceedings of the 8th International Conference on Human-Computer Interaction (HCI International '99) jointly with the 15th Symposium on Human Interface (Japan)*. Fraunhofer IRB Verlag, Stuttgart, 280–281, 1999.

P. Petta, M. Macmahon, A. Staller. FORREST: Forschung ueber/research on emotion simulation. In C. Landauer and K.L. Bellman (eds.), *Proceedings of the Virtual Worlds and Simulation Conference*, 2000 Western Multiconference. Society for Computer Simulation International, San Diego, CA, 2000.

C.M. van Reekum and K.R. Scherer. Levels of processing in emotion-antecedent appraisal. In G. Matthews (ed.), *Cognitive Science Perspectives on Personality and Emotion*. Elsevier, Amsterdam, 259–300, 1997.

I.J. Roseman. Cognitive determinants of emotion: A structural theory. In P. Shaver (ed.), *Review of Personality and Social Psychology* (Vol. 5). Sage, Beverly Hills, CA, 11–36, 1984.

N.J. Saam and A. Harrer. Simulating norms, social inequality, and functional change in artificial societies. *Journal of Artificial Societies and Social Simulation*, 2(1), <http://www.soc.surrey.ac.uk/JASSS/2/1/2.html>, 1999.

C. Saarni. Socialization of emotion. In M. Lewis and J.M. Haviland (eds.), *Handbook of Emotions*. Guilford Press, New York/London, 435-446, 1993.

R.C. Schank. *Dynamic Memory - A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, Cambridge, UK, 1982.

R.C. Schank and R. Abelson. *Scripts, Plans, Goals and Understanding*. Erlbaum, Hillsdale, 1977.

K.R. Scherer. On the nature and function of emotion: A component process approach. In K.R. Scherer and P. Ekman (eds.), *Approaches to Emotion*. Erlbaum, Hillsdale, NJ, 293–318, 1984.

K.R. Scherer. Criteria for emotion-antecedent appraisal: A review. In V. Hamilton, G.H. Bower, and N.H. Frijda (eds.), *Cognitive Perspectives on Emotion and Motivation*. Kluwer, Dordrecht, 89–126, 1988.

K.R. Scherer. Appraisal theory. In T. Dalgleish and M. Power (eds.), *Handbook of Cognition and Emotion*. Wiley, Chichester, 637–663, 1999.

Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies (preliminary report). In *Proceedings of the Tenth National Conference on Artificial Intelligence*. AAAI Press/MIT Press, Cambridge/Menlo Park, 276–281, 1992.

Y. Shoham and M. Tennenholtz. Emergent conventions in multi-agent systems: Initial experimental results and observations (preliminary report). In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*. Kaufman, San Mateo, 225–231, 1992.

C.A. Smith and P.C. Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48:813–838, 1985.

C.A. Smith and L.D. Kirby. Affect and appraisal. In J.P. Forgas (ed.), *Feeling and Thinking: The Role of Affect in Social Cognition*. Cambridge University Press, Cambridge, UK, 2000.

C.A. Smith and R.S. Lazarus. Emotion and adaptation. In L.A. Pervin (ed.), *Handbook of Personality: Theory and Research*, Guilford, New York, 609–637, 1990.

A. Staller and P. Petta. Towards a tractable appraisal-based architecture for situated cognizers. In D. Cañamero, C. Numaoka, and P. Petta (eds.), *Grounding Emotions in Adaptive Systems*, Workshop Notes, 5th International Conference of the Society for Adaptive Behaviour (SAB'98), Zurich, Switzerland, 56–61, 1998.

J.D. Teasdale. Multi-level theories of cognition-emotion relations. In T. Dalgleish and M. Power (eds.), *Handbook of Cognition and Emotion*. Wiley, Chichester, 665–681, 1999.

R.L. Trivers. The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46:35–57, 1971.

A. Walker and M.J. Wooldridge. Understanding the emergence of conventions in multi-agent systems. In *Proceedings of the First International Conference on Multiagent Systems (ICMAS'95)*. AAAI Press, San Francisco, 1995.

M.J. Wooldridge and N.R Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

R.B. Zajonc. Feeling and thinking: Preferences need no inferences. *American Psychologist*, 2:151–176, 1980.

CLIPS Reference Manual: Volumes I&II, Lyndon B. Johnson Space Center, Software Technology Branch, 1993.