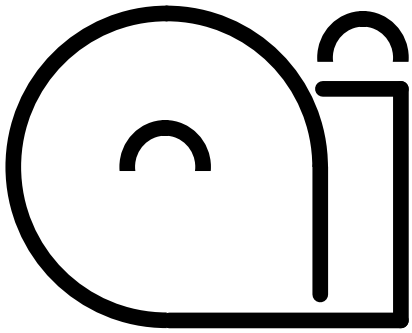


Österreichisches Forschungsinstitut für / Austrian Research Institute for / Artificial Intelligence

TR–2000–27

Alexander Staller, Paolo Petta

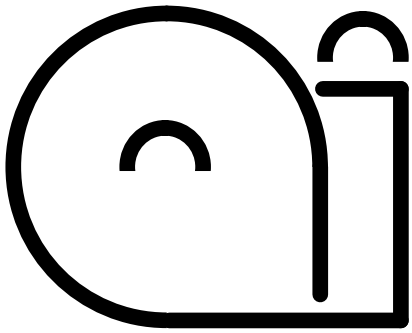
Introducing Emotions into the Computational Study of Social Norms: A First Evaluation

• Schottengasse 3 • A-1010 Vienna • Austria •

• Phone: +43-1-5336112 •

• mailto:sec@ai.univie.ac.at •

• http://www.ai.univie.ac.at/oefai/ •



Österreichisches Forschungsinstitut für / Austrian Research Institute for / Artificial Intelligence

TR–2000–27

Alexander Staller, Paolo Petta

Introducing Emotions into the Computational Study of Social Norms: A First Evaluation

The Austrian Research Institute for Artificial Intelligence is supported by the Federal Ministry of Education, Science and Culture.

Citation: Staller A., Petta P.: Introducing Emotions into the Computational Study of Social Norms: A First Evaluation. In Edmonds B., Dautenhahn K. (eds.): Journal of Artificial Societies and Social Simulation, Special Issue on Starting from Society - the application of social analogies to computational systems, 2001

Introducing Emotions into the Computational Study of Social Norms: A First Evaluation

Alexander Staller and Paolo Petta

Austrian Research Institute for Artificial Intelligence[(*)]
Schottengasse 3, A-1010 Vienna, Austria

To appear in Edmonds B. and Dautenhahn K. (eds.): "Starting from Society --- the Application of Social Analogies to Computational Systems", special issue of the Journal of Artificial Societies and Social Simulation (JASSS)

Abstract

It is now generally recognised that emotions play an important functional role within both individuals and societies, thereby forming an important bond between these two levels of analysis. In particular, there is a bi-directional interrelationship between social norms and emotions, with emotions playing an instrumental role for the sustenance of social norms and social norms being an essential element of regulation in the individual emotional system. This paper lays the foundations for a computational study of this interrelationship, drawing upon the functional appraisal theory of emotions. We describe a first implementation of a situated agent architecture, $TABASCO_{JAM}$, that incorporates a simple appraisal mechanism and report on its evaluation in a well-known scenario for the study of aggression control as a function of a norm, that was suitably extended.

The simulation results reported in the original aggression control study were successfully reproduced, and consistent performances were achieved for extended scenarios with conditional norm obeyance. In conclusion, it is argued that the present effort indicates a promising lane towards the necessary abandonment of logical models for the explanation and simulation of human social behaviour.

Keywords: norms, emotions

1 Introduction: The Interrelation between Social Norms and Emotions

From the classical philosophers onwards it has been a longstanding position that emotions serve no useful function, and in fact have disruptive and disorganising effects on ongoing behaviour, which have to be minimised in order to behave rationally. However, more recently a different view of emotions has emerged. According to this view, emotions are adaptations to the demands of the physical and social environment and thus serve important functions (see Keltner and Gross 1999 for a review of functional accounts of emotions). On the one hand, emotions serve *intrapersonal* functions (Levenson 1999) by coordinating physiological, perceptual, and cognitive processes that enable the organism to respond adaptively to significant environmental challenges and opportunities. On the other hand, emotions serve *social* functions. These developments were complemented by a reassessment of the notion of rationality and its applicability in dynamic open domains (Sousa 1987, Russell and Norvig 1995, Russell 1997).

1.1 Emotions Sustain Social Norms

Keltner and Haidt (1999) review the social functions of emotions at the individual, dyadic, group, and cultural levels of analysis, in particular, the sustenance of social norms. At all levels of analysis it is shown how embarrassment serves an appeasement function, reconciling social relations following transgressions of social norms. It is important that embarrassment also prevents people from violating social norms. Embarrassment is a very unpleasant experience and people forego personal gain to avoid embarrassment (Keltner and Buswell 1997). Embarrassment is by far not the only emotion that serves to sustain social norms. Other so called "self-conscious" emotions such as shame or guilt serve this function, too. Not surprisingly, self-conscious emotions are also called "social" emotions. Tangney (1999) discusses the self-conscious emotions in more detail. It should be mentioned that it is not only the avoidance of negative emotions that causes people to comply with social norms. Forsyth (1994) notes that people often follow norms because they find it personally satisfying.

1.1.1 An Emotion-Based Definition of Social Norms

Not only psychologists investigate social functions of emotions. The economist Jon Elster (1996, 1999) regards emotions as so important for the sustenance of social norms that he includes them in his definition of social norms as injunctions to behaviour with the following features:

- 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."
- 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*.
- 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 would not seem to be a necessary part of a system of social norms, as enforcement of social norms appears to be overdetermined by sanctions *and* emotions. But Elster (1996) argues that emotions are crucial for sanctions to work: A person who is imposing sanctions on the norm violator is herself 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).

1.1.2 Emotions in long-term relationships

An important function of social norms is the restriction of self-interested behaviour in favour of cooperative or altruistic behaviour. The Prisoner's Dilemma is a model of social interactions in which the participants must decide whether to comply with a norm or to pursue self-interest by violating it. Based on the two-person iterated Prisoner's Dilemma, Trivers

(1971) 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.

Frank (1988) proposes that in social scenarios such as the Prisoner's Dilemma some emotions, the so-called "moral sentiments," commit a person to act contrary to immediate self-interest. For example, the predisposition to feel guilt commits a person to cooperate, even if cheating were in her 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. Thereby, emotions such as guilt and anger act as "commitment devices" that alter the material incentives. Even so, 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. For this model to work, 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 the present paper.

1.2 Social Norms Regulate Emotions

Investigating the functions of emotions is important. But even among the proponents of functional accounts of emotions there is no doubt that emotions are not always helpful. Emotions must be *regulated* in order to avoid negative consequences of unrestrained emotional behaviour. An important function of social norms is such a regulation of emotions. In fact, the most popular perspective in the sociology of emotions deals with this function (e.g., Hochschild 1983, Thoits 1990).

This "social constructionist" perspective is based on three postulates:

- Emotions are not irrevocable, biologically-guided, natural phenomena. Rather, they are socially constructed, i.e., they are amenable to social direction, enhancement, and suppression.
- Social construction is accomplished mainly via culturally defined norms that inform individuals about which emotion is suitable in which situation and how it is expressed appropriately. Hochschild (1983) defines the concepts of "feeling rules" and "expression rules." Ekman and Friesen (1975) extensively discuss "display rules" prescribing appropriate expressive behaviour. A large part of emotion socialisation (Saarni 1993) is devoted to learning of these norms.
- Emotions can be managed. A deviant emotional experience can be brought in line with the normative requirement by regulatory processes.

1.3 Structure of the Paper

The aim of this paper is to lay the foundations for the computational study of the interrelation between social norms and emotions, both at the macro level and at the micro level. In section 2 we outline the main components of the emotion process at the micro level, based on the appraisal theory of emotions. This theory allows us to shed new light on an influential study by Conte and Castelfranchi (1995b) in section 3; this study investigated aggression control as a function of a norm. Appraisal theory also builds the theoretical basis of $TABASCO_{JAM}$, an architecture for situated agents aimed at modelling the emotion process. $TABASCO_{JAM}$ is presented in section 4. To perform first simple social simulations with $TABASCO_{JAM}$ agents, we adopt the scenario of the Conte and Castelfranchi (1995b) study. Sections 5 and 6 are devoted to our implementation of the scenario and of $TABASCO_{JAM}$, respectively. In section 7 we present first experimental results. Section 8 concludes the paper.

2 The Emotion Process

The interrelation between social norms and emotions is based on the emotion process at the micro (i.e., intrapersonal) level. For example, how are emotions such as guilt, shame, contempt, and anger elicited? What is the nature of emotional behaviour that brings people to appease after a norm violation or to punish norm violators? How do regulatory processes work? Relying on Frijda (1986,1995), the following is a description of the main components of the emotion process.

2.1 Appraisal

Emotions are a key element in successful coping with a non-deterministic, dynamic, and social environment. This coping depends on the continuous monitoring of the relationship between the individual and the environment. The (largely unconscious) cognitive process underlying this monitoring is called *appraisal*.

2.1.1
Appraisal has attracted much attention by psychologists. In fact, the "appraisal theory of emotions" (see Scherer 1999 for a review) has become the predominant approach to psychological research on emotions. The central tenet of appraisal theory "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.

2.1.2
Many theorists have been trying to specify the criteria according to which a situation is appraised (e.g., Scherer 1984; Smith and Ellsworth 1985; Frijda 1986; Ortony, Clore and Collins 1988; Lazarus 1991; Roseman, Antoniou and Jose 1996). 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."

2.1.3
The above characterisation of the appraisal criteria contains the term *concerns*. Frijda (1986, p. 335) defines a concern "as a disposition to desire occurrence or nonoccurrence of a given kind of situation." Humans have many concerns, e.g., the biological ones such as concerns for the optimal state of feeding, drinking, and temperature, as well as concerns for being well-oriented and for the proximity of trusted individuals ("attachment figures"). Especially the biological concerns can be thought of as *setpoints* representing a desired state (e.g., a specific glucose level). If the mismatch between the actual circumstances and the setpoint exceeds a certain *threshold*, the concern becomes unsatisfied. Then stimuli suitable to reduce the mismatch (e.g., food) are appraised as relevant for concern satisfaction, until the setpoint is reached again. There are other ways in which stimuli can be appraised as relevant, e.g., as threats or obstacles to concern satisfaction. In addition to the personal concerns mentioned so far, culturally defined *values* can become concerns of humans. Justice, honour, and the conformation to social norms are examples of such values.

2.1.4
Social norms enter the process of emotion generation during appraisal. 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" (Frijda 1986, p. 311). This list makes clear that to differentiate these emotions, the appraisal criterion "agency or responsibility" is required. 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 (see also e.g. Ortony et al. 1988; Roseman et al. 1996).

2.1.5
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) criticised 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. Currently, there is a trend towards multi-level theories of the appraisal process. For example, Smith and Kirby (2000) suggest a model of the appraisal process in which perceptual processing is complemented by associative processing and reasoning. 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.

2.2 Impulse

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" (Frijda 1986, p. 75). In addition to these "relational" action tendencies, Frijda (1986) identifies two other kinds of

action tendencies: the tendency to engage in consummatory behaviour (e.g., eating) and the tendency to approach or bring about situations of satisfaction (e.g., to approach food). An impulse is best understood as a *goal* which can be achieved by different *plans* (Frijda 1995). For example, the goal to remove an obstacle for concern satisfaction (as in anger) can be achieved by different forms of aggressive behaviour, such as a bodily attack or a verbal threat; Frijda (1986) calls the action tendency underlying aggressive behaviour "agonistic." Further, an impulse has the feature of *control precedence*: It tends to interrupt ongoing processes and to take control over behaviour, attention and resources. Simon (1967) already ascribed the function of an interrupt mechanism to emotions.

2.3 Action

Action tendencies eventually lead to action. Frijda (1986) distinguishes between instrumental activity and expressive behaviour. In contrast to expressive behaviour such as facial expressions, instrumental activity directly changes the worlds objective state by overt action (e.g., by attacking a rival) or by cognitive action (e.g., by deprecating a rivals worth in ones mind).

2.4 Regulation

All steps of the emotion process sketched so far are subject to regulatory processes. These 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. Emotion regulation requires *anticipation of the consequences* of emotional responses prior to execution, either based on memory or on the computation of consequences when the response is still in the planning stage. Consequences of emotional responses can be external, e.g., retaliation following displayed anger, or internal, such as the interruption of ongoing nonemotional task performance.

With respect to potential norm violations through unrestrained emotional behaviour, there are also external and internal consequences that must be anticipated by the regulatory processes. An external consequence is punishment through sanctions. An internal consequence is the generation of an emotion such as shame or guilt contingent upon the norm violation.

2.5 Emotion Intensity

One of the most noticeable aspects of an emotion is its intensity. Emotion intensity is a very complex phenomenon in two respects: First, intensity is not unidimensional. There are many parameters that can vary in magnitude, e.g., the duration of an emotion, the delay of its onset, and the strength and drasticness of action tendency (Frijda et al. 1992; Sonnemans and Frijda 1994). Second, there are many determinants of intensity, e.g., concern strength (Sonnemans and Frijda 1995) and cognitive determinants such as the appraisal criteria of praiseworthiness, desirability, and appealingness (Ortony et al. 1988; Melton et al. 1993).

A special case has been investigated thoroughly, namely the influence of concern strength on the intensity of aggressive behaviour. The so-called "frustration-aggression hypothesis" states that the stronger the concern whose satisfaction is obstructed, the more drastic is the agonistic action tendency aimed at removing the obstruction (Frijda 1986).

3 Aggression Control as a Function of a Norm

3.1 An Influential Study

Conte and Castelfranchi (1995b) realized that previous work in Artificial Intelligence (Shoham and Tennenholtz 1992a, 1992b) 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 (1995b) 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, Conte and Paolucci 1998; Saam and Harrer 1999).

In the study by Conte and Castelfranchi (1995b), agents move through a two-dimensional world with randomly scattered food items. The world is a 10 x 10 grid with connected edges (a torus). Each cell of the grid has the capacity of holding one agent and one food item simultaneously. Agents move through the world in a discrete fashion. In one step an agent can move one cell up, down, left, or right, but not diagonally. Agents move around and stop to eat when they are on a cell with a food item. Eating agents can be attacked by agents located at neighbouring cells. The agents' sensing capabilities consist of seeing and smelling. An agent can "see" food items and agents within its "territory," consisting of the four neighbouring cells above, below, to the left, and to the right. Food can be smelt within the agent's "horizon," consisting of the eight cells which can be reached with two movement steps.

A simulation step is called a "game." In each game, all agents except those eating select an action from the following routines (listed in the order of preference): EAT, MOVE-TO-FOOD-SEEN, MOVE-TO-FOOD-SMELT, AGGRESS, MOVE-RANDOM, and STAY. If an agent is located at a cell with food, it chooses EAT. Eating takes two games, unless it is interrupted by aggression. When eating has been completed, the eater's strength is increased by the food's nutritional value and the food item is restored at a randomly chosen cell. The selections following in the order of preference are MOVE-TO-FOOD-SEEN and MOVE-TO-FOOD-SMELT. The precondition for choosing MOVE-TO-FOOD-SEEN is that the agent sees a cell with an unoccupied food item within its territory. MOVE-TO-FOOD-SMELT is chosen if the agent smells food items within its horizon. If none of the above choices is possible, the agent checks whether an eating agent is located at a neighbouring cell. Depending on a strategy for aggression detailed below, it may choose to AGGRESS the eating agent. In an aggression, the stronger agent is the winner and obtains the food item. If the competitors are equally strong, the attacked agent keeps the food item. The selections lowest in the order of preference are to MOVE-RANDOM to a free neighbouring cell and to STAY (if no free neighbouring cell is found). Actions are costly: Moving reduces the agent's strength by 1; the cost of attacking or being attacked is 4; the cost for STAY is 0.

Actions are supposed to be executed simultaneously. So conflicts can arise: When several agents choose to move to the same cell, one agent is given preference at random while the other agents STAY at their locations. When several agents choose to attack the same agent, the strongest aggressor receives the food item, provided that its strength is higher than the attacked agent's strength. If more than one aggressor share the highest strength, one is chosen at random. The cost for the attacked agent is multiplied by the number of aggressors.

Conte and Castelfranchi (1995b) investigated the role of norms in the control of aggression. To this end, three experimental conditions were designed:

1. *Blind* aggression: Agents attack eaters without taking their own or the eaters' strength into account.
2. *Strategic* aggression: Agents can detect the strengths of eaters within their territory. They only attack eaters whose strength is not higher than their own.
3. *Normative* aggression: At the beginning of a match, agents are "assigned" the food items falling into their territories, i.e., agents become "finders" of these food items. A norm of precedence to finders is introduced whereby finders become possessors of food. Agents can determine to whom a food item belongs and will not attack possessors eating their own food.

For each condition an experiment consisting of 100 matches, each comprising 2000 games, was conducted. At the beginning of a match, 50 agents and 25 food items of nutritional value 20 are placed randomly on the grid. The initial strength of all agents is set to 40. For each experiment, the number of attacks, the average strength, and the standard deviation of average strength (as a measure of inequality) was recorded, and the significance of the differences tested. The normative strategy has been found to do best at constraining aggression, at promoting average strength, and at keeping inequality low.

3.2 An Emotion-Based View

In appraisal theory terminology, aggressive behaviour is a prominent example of emotional behaviour that stems from the agonistic action tendency and has the goal to remove an obstacle for concern satisfaction. In this light, a considerable part of computational research on social norms has actually investigated an instance of emotion regulation through social norms, namely the control of aggression through the "finder-keeper" norm. However, neither Conte and Castelfranchi (1995b) nor the authors of the follow-up studies mentioned emotions.

Conte and Castelfranchi (1995b) 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. Our goal is to investigate both macro-level and micro-level aspects of the interrelation between social norms and emotions. To this end, we adopted the scenario of the Conte and Castelfranchi (1995b) study, except for the simple action selection algorithm of the agents. To study the micro-level processes underlying the interrelation between social norms and emotions, social simulations with more complex agents whose architecture models the emotion process must be conducted. Such an architecture should also model behaviour of living beings more accurately with respect to the following two points: First, animals are not constantly searching for food and eating. Only at the instigation of internal signals they interrupt their ongoing activities to search for food and eat. When they are satiated, they stop eating and continue their activities. Second, humans are capable of obeying and violating the same norm on different occasions. The agents in the Conte and Castelfranchi (1995b) study either rigidly obey or violate the norm throughout the whole simulation. The agent architecture presented in the next section models the emotion process and remedies these shortcomings.

4 Towards an Agent Architecture Modelling the Emotion Process

4.1 The TABASCO Project

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; Elliott 1997), Tok (Bates, Loyall and Reilly 1992), or FLAME (Seif El-Nasr, Yen and Ioerger 2000). The reification of emotional states makes it necessary to implement an explicit mapping from these states to entailed effects, including internal processing and externally observable overt behaviour. Problems of systems engineered according to such a shallow approach are well known from the traditional research area of expert systems in artificial intelligence: brittleness of system behaviour surfacing with every occurrence of any situation not explicitly anticipated at implementation time; laboriousness of system extension; and difficulties in maintaining consistency with a growing body of incorporated knowledge. Besides these problems, 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.

In the light of these facts we have taken a different approach for endowing agents with emotions with TABASCO (a **T**ractable **A**ppraisal-**B**ased **A**rchitecture for **S**ituated **Co**gnizers), an architecture for agents situated in virtual environments. In TABASCO, emotions are not modelled as reified entities. Instead, we aim at capturing the main components of the emotion *process* (section 2), which is seen as a crucial adaptive element of the agent-environment interaction. Staller and Petta (1998) provide some theoretical background for the development of TABASCO.

In trying to flesh out TABASCO, we have been examining established agent architectures for situated agents, analyzing and relating their characteristics to our approach of modeling the emotion process. In particular, various papers such as (Horswill 1997) and (Gat 1997) make a strong point in favour of a layered approaches that instantiate “useful intermediate points” between “traditional and reactive systems”. Well known examples include $_3$T (Bonasso et al. 1997) and the family of systems derived from the Procedural Reasoning System (PRS) (Georgeff and Lansky 1987). $_3$T is a three-layer architecture integrating reactive behaviour with a traditional planner. This integration is accomplished by the RAP (Reactive Action Packages) system (Firby 1989) for the reactive execution of plans. PRS is another approach for integrating goal-directed reasoning and reactive behaviour. PRS is a so-called BDI architecture because it is built on the philosophical tradition of understanding *practical reasoning* based on the concepts of *beliefs*, *desires*, and *intentions*. There are several implementations of PRS, including UM-PRS (Lee et al. 1994) and PRS-CL (Myers 1997). Following this empirical evidence, we followed a similar layered approach in first successful implementations situated in text-based (Petta, Macmahon and Staller 2000) and immersive graphical interactive virtual environments (Petta 1999; Petta et al. 1999).

In the following section we first describe JAM, an agent architecture that draws upon the theories and ideas of PRS and on the implementation pragmatics of UM-PRS and PRS-CL. Then we will present how the main components of the emotion process can be modelled on top of JAM, making it a suitable starting point for furthering our research in the TABASCO project.

4.2 The JAM Architecture

Huber (1999a, 1999b) provides a detailed description of the JAM architecture, which is implemented in Java. A JAM agent consists of five components (see Figure 1): a *World Model*, a *Plan Library*, an *Interpreter*, an *Intention Structure*, and an *Observer*. The World Model is a database representing the *beliefs* of the agent. The Plan Library is a collection of

plans the agent can use to achieve its goals (i.e., its *desires*), which are specified by the agent developer. The Interpreter is responsible for selecting and executing plans. The Interpreter is associated with the Intention Structure, a run-time stack of goals with and without instantiated plans. Instantiated plans contained in the Intention Structure are the agent's *intentions*. The Observer is a lightweight declarative procedure that the agent interleaves between plan steps.

The Interpreter operates in a loop. In each cycle of the loop, it computes the *Applicable Plan List* (APL), consisting of the instances of all plans specified in the Plan Library that can be applied to the current goals. The computation of the APL is based on the beliefs in the World Model. Plans and goals have utility values. The utility of a plan instance in the APL is the sum of the utilities of the plan and the goal to which it is applicable. Based on *metalevel plans* or on maximum utility, the Interpreter decides to intend an element of the APL (i.e., commits itself to execution of an instantiated plan) and subsequently executes the next action of the plan associated with the intention of the highest current utility. The Observer procedure is executed at the beginning of each cycle. Typically, the Observer watches for asynchronous events and is used for sensing.

The agent's behaviour depends on its *top-level goals*. Top-level goals are persistent: they are given to the agent when it is started and pursued until they are satisfied or removed explicitly within a plan. Goals can not only be given to the agent at invocation. It is also possible to post goals to the Intention Structure during run-time.

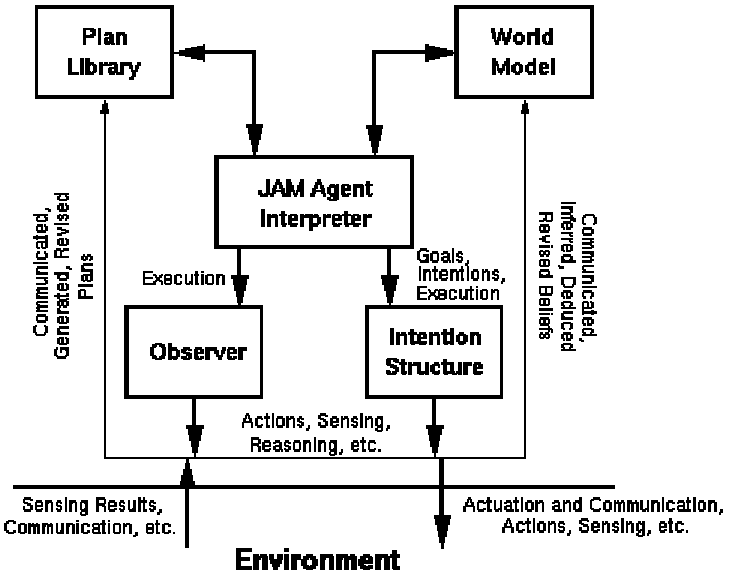


Figure 1: The JAM architecture (adapted from Huber 1999a)

4.3 The TABASCO$_{JAM}$ Architecture

In the following, we outline how the main components of the emotion process can be modelled on top of JAM (see Figure 2):

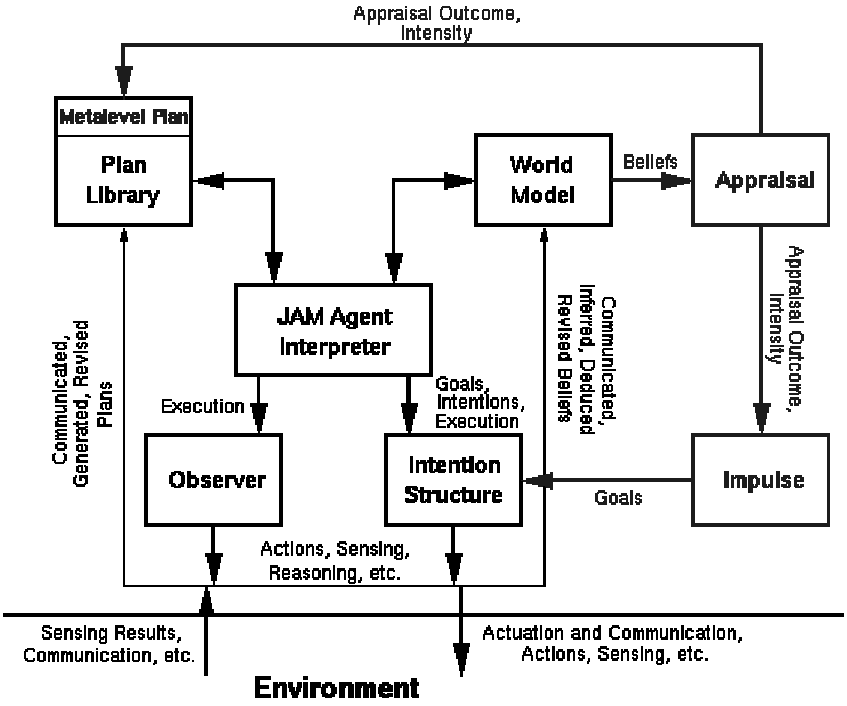


Figure 2: The $TABASCO_{JAM}$ architecture

- **Sensing:**
  The Observer senses the world and asserts the sensing results to the World Model.
- **Appraisal:**
  An Appraisal component is added to JAM. The Appraisal component models the appraisal process and maps beliefs contained in the World Model (e.g., about sensing results and internal states of the agent) to the appraisal outcome, the construed "situational meaning structure" (Frijda 1986) framing the appraisal criteria (e.g., relevance: "concern satisfaction obstructed", agency: "other", coping potential: "high"). Depending on the complexity of the environment and the sensing data, this mapping can be very complex. Unfortunately, at the moment there is no detailed theory of the appraisal process. Multilevel theories (2.1.5) can at least provide a guideline. The appraisal of concern relevance can be based on the conception of concerns as setpoints (2.1.3).

  Currently, emotion intensity is modelled as a numerical value. The Appraisal component computes an intensity value, e.g., based on concern strength and on appraisal criteria hypothesised as cognitive determinants of emotion intensity (2.5).

- **Impulse:**
  A second component added to JAM is the Impulse component. It receives the appraisal outcome and the intensity value from the Appraisal component. Based on the conception of an impulse as a goal in connection with plans that are put in a state of readiness (2.2), JAM is well suited to model the generation of action tendencies. Depending on the appraisal outcome (e.g., indicating that concern satisfaction is obstructed), the Impulse component posts a goal to the Intention Structure (e.g., the goal to remove the obstruction). The utility of the posted goal is currently taken to be the intensity value calculated by the Appraisal component. At present it is assumed that the Plan Library contains at least one plan that is applicable to the posted goal. The computation of the APL is equivalent to putting plans in a state of readiness, i.e., to the generation of action tendencies. The control precedence feature of impulses

results from JAM's built-in mode of operation: If a plan instance in the APL is intended and the utility of the new intention is the highest among all intentions, the Interpreter interrupts the execution of the previous intention and executes the new one. Thus, the higher the intensity value, the more likely is an interruption of ongoing processes.

- **Action:**
  The plans in the Plan Library that are applicable to the goals posted by the Impulse component contain the actions to be executed (e.g., attacking a rival).
- **Regulation:**
  Currently we do not intend to model the multitude of regulatory processes, but concentrate on modelling impulse control. The regulatory process responsible for impulse control can be modelled as a metalevel plan that decides whether a plan instance in the APL applicable to a goal posted by the Impulse component (i.e., an action tendency) is intended or not. The interesting case that the metalevel plan relies on a social emotion for this decision can be modelled in the following way:
  1. The metalevel plan asserts a fact to the World Model specifying the plan instance in the APL under consideration.
  2. The Appraisal component reads the asserted fact and determines whether the execution of this plan instance results in a norm violation. If so, it computes an intensity value.
  3. The metalevel plan receives the appraisal outcome and the intensity value from the Appraisal component and subtracts the intensity value from the utility of the plan instance to be evaluated.
  4. If the resulting utility is negative, the plan instance is excluded from further metalevel reasoning. This means that the plan instance will not be intended; the metalevel plan decides to obey the norm.
  5. Otherwise, the metalevel plan continues to consider the plan instance, but with the reduced utility value. Eventually, the metalevel plan may decide to intend the plan instance. This means that the metalevel plan decides to violate the norm.

  Social norms are not only directed towards emotion regulation but to the regulation of behaviour in general. The above algorithm is not restricted to plan instances applicable to goals posted by the Impulse component, but to plan instances specifying arbitrary behaviour. It is not only a way of modelling the role of social emotions for emotion regulation through social norms, but for behaviour regulation in general.

- **Task-directed behaviour:**
  Emotion processes interrupt task-directed processes and arise as consequences of events encountered during task execution or as consequences of events external to the task at hand. The specification of task-directed behaviour is the normal use of JAM and is simply done by specifying toplevel goals and plans for accomplishing them.

All in all, JAM seems to provide a suitable basis for our research work, allowing us to focus on the innovative aspects of the realization of tractable appraisal-based agent control systems. As a powerful BDI architecture, JAM frees us of much tedious implementation work without overly constraining the space of accessible designs. In addition, the ease of integration of additional Java code provides further flexibility.

5 Implementation of the Scenario

For deployment of situated agents, a suitable environment must be provided. As already discussed, in the present study we adopt the scenario of the Conte and Castelfranchi (1995b) study (3.1). The scenario is implemented such that TABASCO$_{JAM}$ agents running as independent processes on different computers can participate in social simulations. Figure 3 shows the set-up. The World maintains information about each cell of the grid (e.g., presence of food and agents) and each agent (e.g., position and strength). As the scenario prescribes synchronous execution of discrete "games" where the actions selected by participating agents are assumed to be executed simultaneously, the agents are not provided with direct access to the world. Instead, they interact with a central Controller via socket connections. The Controller runs the loop for the matches and games. At the onset of each match the World computes random locations of the agents and food items and determines the owners of the food items. In each game, the Controller gets the sensing results for all agents from the World and sends them to the agents (except for eating agents which must pause). Then, it receives the selected action from each agent. Only after the Controller has collected the actions of all agents, the World simulates the synchronous execution of the actions. The assumptions about the restoration of eaten food items, the costs of actions, and the resolution of conflicts are adopted from Conte and Castelfranchi (1995b).

The sensing results comprise information about the current strength, the presence of food at the location, the presence of food and agents in the territory, the presence of food within the horizon, the strengths of the agents in the territory, and the ownership of food being eaten in the territory. The possible actions that can be selected by the agents and sent to the Controller are: EAT, MOVE UP | DOWN | LEFT | RIGHT, AGGRESS UP | DOWN | LEFT | RIGHT, and STAY.

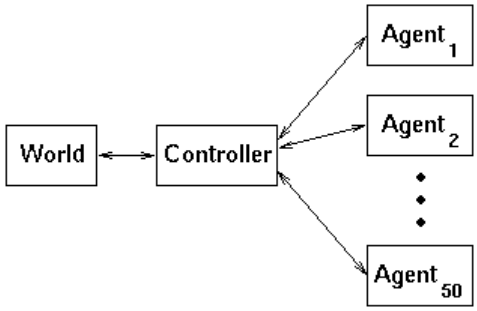


Figure 3: The set-up of the scenario implementation

6 Implementation of TABASCO$_{JAM}$

6.1 Specification of the Scenario-Specific Micro-Level Processes

The first step towards an implementation of TABASCO$_{JAM}$ is the specification of the scenario-specific micro-level processes to be modelled:

- **Sensing:**
  The pertinent information about food and other agents must be sensed.
- **Appraisal:**
  To simplify matters, only the appraisal of concern relevance is considered: The optimal state of feeding is a basic concern of living beings. As long as the concern is not satisfied, food is appraised as relevant for concern satisfaction and an eating agent

is appraised as obstructing concern satisfaction. If the concern is satisfied, neither food nor eating agents are appraised as relevant. A second concern is the compliance with the "finder-keeper" norm. Attacking an agent eating its own food is appraised as concern relevant because the norm is violated. Observations of attacks are not considered.

- **Impulse:**
  The appraisal of food as concern relevant leads to the generation of the tendencies to eat (i.e., to execute consummatory behaviour) or to approach food (i.e., to approach or bring about a situation of satisfaction). The appraisal of an eating agent as obstructing concern satisfaction leads to the generation of the agonistic action tendency with the goal to remove the obstruction.
- **Action:**
  Eating, moving towards seen or smelt food, and attacking are instrumental activities stemming from the above action tendencies.
- **Regulation:**
  Aggression control is an instance of impulse control. The regulatory process must be able to prevent the agonistic action tendency from leading to overt aggressive behaviour. Especially the case that the regulatory process relies on a social emotion contingent upon a norm violation must be modelled.
- **Task-directed behaviour:**
  Some task-directed behaviour must be modelled that is executed when the concern for the optimal state of feeding is satisfied.

6.2 Implementation Details

In the following we describe the implementation of the above processes in more detail:

- **Sensing:**
  The Observer establishes a socket connection with the Controller when the agent is started. After an action has been sent to the Controller by one of the executed plans, the Observer receives the new sensing results from the Controller and uses them to update the World Model.
- **Appraisal:**
  Currently only the appraisal of concern relevance is modelled. Based on information read from the World model, the Appraisal component assigns a specific value to the variable `relevance`. Additionally, the direction of the appraised event ("UP", "DOWN", "LEFT", or "RIGHT") is assigned to the variable `direction`. The Appraisal component also computes an intensity value which is assigned to the variable `intensity`. The variables `relevance`, `direction`, and `intensity` are output variables accessible from other system components. The Appraisal component contains two objects, one of the class `OptimalFeedingConcern` and one of the class `NormComplianceConcern`. Each of these objects contains a method mapping information from the World Model to specific values of the output variables. Before this mapping is performed, the output variables are initialised as follows: `relevance` = "", `direction` = "", and `intensity` = 0.
    1. `OptimalFeedingConcern`:
        - Based on Frijda's conception of concerns as setpoints in connection with thresholds (2.1.3), the class contains two parameters: `feedSetpoint` > 0 and `feedThreshold` >= 0.
        - Then it is checked whether the concern is satisfied or not: `feedConcernStrength := feedSetpoint - agentStrength`

If `feedConcernStrength > feedThreshold`, the concern is unsatisfied. It remains unsatisfied until `agentStrength >= feedSetpoint`.

- If the concern is unsatisfied, the output variables are assigned values depending on sensing results contained in the World Model, as shown in Table 1:

Table 1

World Model	**Output variables**
agent's location provided with food	`relevance`: "object for concern satisfaction reached" `intensity`: `feedConcernStrength` + 3 `direction`: ""
free food is seen up (down, left, right)	`relevance`: "object for concern satisfaction seen" `intensity`: `feedConcernStrength` + 2 `direction`: "UP" ("DOWN", "LEFT", "RIGHT")
food is smelt up (down, left, right)	`relevance`: "object for concern satisfaction smelt" `intensity`: `feedConcernStrength` + 1 `direction`: "UP" ("DOWN", "LEFT", "RIGHT")
eating agent is seen up (down, left, right)	`relevance`: "concern satisfaction obstructed" `intensity`: `feedConcernStrength` `direction`: "UP" ("DOWN", "LEFT", "RIGHT")

- Concern strength is an important determinant of emotion intensity in general and of the drasticness of the agonistic action tendency in particular (2.5). We chose to take `feedConcernStrength` directly as the intensity value, in connection with additive offsets. There are two reasons for the offsets: A practical reason is that we want to keep the order of preference that has been chosen by Conte and Castelfranchi (1995b). This facilitates a comparison between the behaviour of the $TABASCO_{JAM}$ agents and the agents of the original study. A theoretical reason is that proximity is also an important determinant of emotion intensity (Frijda 1986). For example, food that is found at the location instigates a more intense reaction than food that is seen or smelt.

2. `NormComplianceConcern`
   - This class contains the parameter `normConcernStrength` indicating the strength of the concern for compliance with the "finder-keeper" norm.
   - The metalevel plan responsible for impulse control (see below) relies on this class. After the metalevel plan has asserted a fact to the World

Model indicating that a plan instance to perform the agonistic action tendency is under consideration, a method reads this fact and determines whether the eater to be attacked is the owner of the food item (this information is a sensing result contained in the World Model). If so, the output variables are set in the following way:
`relevance`: "norm violated"
`intensity`: `normConcernStrength`
`direction`: <direction included in the fact asserted by the metalevel plan>

- Note: Concern strength is again taken directly as the intensity value.

- **Impulse:**
  Depending on the values of the variables `relevance`, `direction`, and `intensity`, the Impulse component posts goals to the Intention Structure as shown in Table 2. The value of `intensity` is taken as the utility of the posted goal.

Table 2

Value of `relevance`	**Posted goal**
"object for concern satisfaction reached"	PERFORM action_tendency "consummatory"
"object for concern satisfaction seen"	PERFORM action_tendency "approach_seen_food" <value of `direction`>
"object for concern satisfaction smelt"	PERFORM action_tendency "approach_smelt_food" <value of `direction`>
"concern satisfaction obstructed"	PERFORM action_tendency "agonistic" <value of `direction`>

- **Action:**
  One plan is applicable to each posted goal. The actions sent to the Controller are included in the plans. Table 3 shows the posted goals on the left and the actions on the right:

Table 3

Posted Goal	**Action**
PERFORM action_tendency "consummatory"	EAT
PERFORM action_tendency "approach_seen_food" <value of `direction`>	MOVE <value of `direction`>
PERFORM action_tendency "approach_smelt_food" <value of `direction`>	MOVE <value of `direction`>
PERFORM action_tendency "agonistic" <value of `direction`>	AGGRESS <value of `direction`>

- **Regulation:**
  The metalevel plan responsible for impulse control looks as follows:
  1. For each plan instance in the APL:
     - If the goal of the plan instance is to perform the agonistic action tendency:
       - Assert a fact indicating the plan instance to the World Model.
       - Subtract the intensity value computed by the object of the class `NormComplianceConcern` from the plan instance's utility.
       - Retract the asserted fact from the World Model.
  2. From all APL elements with a utility >= 0, intend the APL element with the highest utility. If more APL elements share the highest utility, select one randomly and intend it.
  3. Remove the goals of all unintended APL elements from the Intention Structure (this step is necessary because in the following game the plan instances are not applicable any more due to new sensing results)
- **Task-directed behaviour:**
  As task-directed behaviour the agent moves randomly around to free cells (i.e., cells that are not provided with food and are not locations of other agents). If no such cell is found the agent stays at its location. Since the cost of moving is 1, an agent's strength gradually decreases by performing this behaviour, so that the concern for the optimal state of feeding may become unsatisfied again (depending on the values of `feedSetpoint` and `feedThreshold`).

  This behaviour is implemented by specifying a toplevel goal that remains on the Intention Structure throughout the whole simulation. The plan applicable to this goal is a loop. At each cycle of the loop, the agent randomly selects a free neighbouring cell and sends the appropriate MOVE action to the controller. If no free cell is found, the action STAY is sent. The utilities of the goal and the plan are 0. The Impulse component only posts goals to the Intention Structure, if `feedConcernStrength` > `feedThreshold` >= 0. `feedConcernStrength` determines the utility of the posted goals. So the utilities of goals posted by the Impulse component are always greater than 0, ensuring that task-oriented behaviour is interrupted by the Interpreter in favour of emotional behaviour.

7 Experiments

7.1 Experiment 1

First we tried to replicate the results of the Conte and Castelfranchi (1995b) study. To this end, we developed an agent that is able to interact with the Controller, and uses the original action selection algorithm. In the order of preference, it chooses EAT, MOVE <direction in which food is seen>, MOVE <direction in which food is smelt>, AGGRESS <direction in which eating agent is seen>, MOVE <direction in which free cell is seen>, or STAY. In case there are more directions to choose from, one is selected randomly. In two of the three experimental conditions, the agent additionally checks a precondition that must be fulfilled to select an AGGRESS action. In the *blind* condition, there is no additional precondition; in the *strategic* condition the eating agent must not be stronger than the agent; in the *normative* condition the eating agent must not be the owner of the food item.

In each condition 100 matches, each consisting of 2000 games, were run. After each match the following measures were recorded: the number of aggressions (**Agg**), the average strength of the agents (**Str**), and its standard deviation (**Dev**). Table 4 shows the means and standard deviations of these measures computed over the 100 matches. Figure 4 depicts the means. The significance of the differences between the means was tested with an ANOVA. All differences are significant ($p < .001$).

The proportions between the lengths of the bars in Figure 4 look very much like in the graph of the original study (Conte and Castelfranchi 1995b, p. 261, Fig. 13.4). However, the absolute numbers of our results are much higher. We do not know the reason for this discrepancy. However, Castelfranchi et al. (1998) replicated the original study themselves and received absolute numbers much closer to our results, but do not provide any reason for the discrepancy between their results and the results of the Conte and Castelfranchi (1995b) study. The results obtained by Castelfranchi et al. (1998) are shown in Table 5.

In sum, we were successful in replicating the major results. The normative strategy does best at constraining aggression, at promoting average strength, and at keeping inequality low; the blind strategy does worst at constraining aggression and promoting average strength; and the deviation of strength for the strategic strategy is higher than for the blind strategy.

Table 4: Results of experiment 1

	Agg	st. dev.	**Str**	st. dev.	**Dev**	st. dev.
Blind	9245	710	4027	89	1723	112
Strategic	4684	231	4656	35	1823	75
Normative	1474	59	5958	19	418	28

Table 5: Results of the replication experiment by Castelfranchi et al. (1998)

	Agg	st. dev.	**Str**	st. dev.	**Dev**	st. dev.
Blind	9235	661	4287	204	1443	58
Strategic	4634	248	4727	135	1775	59
Normative	3018	76	5585	27	604	41

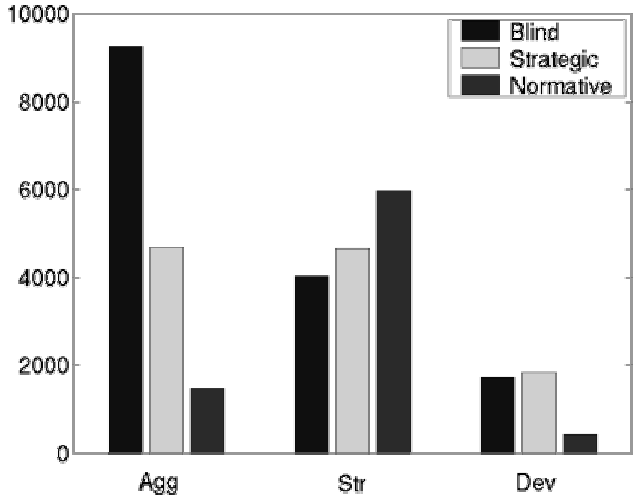


Figure 4: The means of the number of aggressions (**Agg**), the average strength (**Str**), and the deviation of strength (**Dev**) for the three conditions

7.2 Experiment 2

The aim of this experiment was to study the effects of the agents' ability to decide whether to obey or violate the "finder-keeper" norm, based on the strengths of their concerns for the optimal state of feeding and for norm compliance. Three parameters can be chosen: `feedSetpoint`, `feedThreshold`, and `normConcernStrength`. We wanted to compare the results of this experiment with the results of the replication experiment, in which the agents are constantly in search of food and eating. So `feedSetpoint` was chosen so high that it cannot be reached by the agents in 2000 games. We chose the following parameter settings: `feedSetpoint` = 8000 and `feedThreshold` = 0. These parameter settings ensure that the concern for the optimal state of feeding is always unsatisfied for agents with an initial strength of 40.

The parameter `normConcernStrength` determines whether an agent decides to violate the norm. For example, consider an agent with a current strength of 2500. The strength of the concern for the optimal state of feeding is `feedSetpoint` - 2500 = 8000 – 2500 = 5500. Thus the utility of a plan to attack an eater is 5500. If the eater is the owner of the food, the metalevel plan subtracts `normConcernStrength` from this utility. So if `normConcernStrength` is, say, 6000, the reduced utility is negative (-500) and the plan to attack is abandoned. On the other hand, if `normConcernStrength` = 5000, the reduced utility is positive (500). If no other plan instances with higher utility are in the APL (e.g., plans to eat or to move to food), the attack will be executed. An agent with `normConcernStrength` = `feedSetpoint` always obeys the norm, since the strength of the feeding concern can never be higher than `feedSetpoint`. On the other hand, an agent with `normConcernStrength` = 0 always violates the norm, since the utility of a plan to attack is never reduced and thus the plan is never abandoned. This choice of range for `normConcernStrength` thus includes Conte and Castelfranchi's blind and normative agents at the ends of the scale.

In the experiment, `normConcernStrength` varied from 0 to `feedSetpoint` (8000) in steps of 1000. The same measures as in experiment 1 were computed. Table 6 shows the results for the different values of `normConcernStrength`. Figures 5, 6, and 7 depict **Agg**, **Str** and **Dev**, respectively. As expected, the results of experiment 1 were replicated: for blind aggression with a value for `normConcernStrength` of 0, and for normative aggression with a value for

normConcernStrength equal to the feedSetpoint, respectively (the differences in the reported numbers lie below the threshold of statistical significance). For these three measures, the significance of the differences between succeeding means (as ordered by normConcernStrength) was also tested. For **Agg** all differences of pairs except for the first two (i.e., between **Agg** (normConcernStrength) values of 9308(0), 9281(1000) and 9199(2000)) are significant ($p < .05$). For **Str** all differences but for the first (for normConcernStrength values of 0 and 1000) are significant ($p < .01$). In contrast, for **Dev** only every other difference (between pairs with normConcernStrength values of 1000 and 2000, 3000 and 4000, 5000 and 6000, 7000 and 8000) is significant ($p < .05$).

The performance with respect to all three measures improves with increasing values of normConcernStrength. Agents always violating the norm (normConcernStrength = 0) perform worst, while agents always obeying the norm (normConcernStrength = feedSetpoint (8000)) perform best. As can be seen in Figure 5, **Agg** values remain virtually constant for normConcernStrength values up to 2000, and then start to decrease until reaching an approximately linear falling slope for normConcernStrength values of 5000 and higher. A similar pattern can be observed for **Str** (see Figure 6), where the slope of increase roughly stabilises for normConcernStrength values of 4000 and higher.

The pattern for **Dev** (see Figure 7) differs considerably from the patterns for **Agg** and **Str**: no consistent simple pattern can be observed for lower normConcernStrength values, and there is a marked drop between normConcernStrength values of 7000 and 8000.

The means of the number of aggressions for small normConcernStrengths cannot be expected to differ greatly, as the agents first have to reach a sufficient strength (of feedSetpoint - normConcernStrength) during the simulation so as to respect the norm. As the threshold value decreases (with increasing normConcernStrength), more agents start to respect the norm earlier on in the simulation, leading to increasingly smaller average numbers of aggressions. This explanation holds analogously for the development of the means of the average strength over increasing values of normConcernStrength.

Figures 8, 9 and 10 plot the development of **Agg**, **Str**, and **Dev** in the course of the experiments. As the norm does not apply to all kinds of aggression scenarios, it only reduces the frequency of occurrences of situations in which an agent will aggress another. Figure 8 shows how this dampening effect comes about the earlier the sooner the agents start obeying the norm, leading to the transition from the steeper slope characteristic of norm disregard to the flatter slope for agent populations in which the norm is obeyed. Conversely, Figure 9 shows a flatter slope for norm disobeyance and a steeper one for norm obeyance. The transitions between these two slopes occur at the number of games at which **Str** reaches the value of feedSetpoint - normConcernStrength (Figure 9). For example, for normConcernStrength = 6000, **Str** reaches this value (feedSetpoint (8000) - 6000 = 2000) after approx. 1000 games (Figure 9). Accordingly, the **Agg** plot for normConcernStrength = 6000 flattens after 1000 games (Figure 8). Conversely, a small increase in the gain of **Str** can be seen in Figure 9.

The setup of the scenario under investigation favours a general increase in strength (Figure 9) with an increasing spread between stronger and weaker agents over time (Figure 10). Further, the norm does not differentiate between stronger and weaker agents (it is not a function of an agent's strength). It follows that obeyance of the norm has only a limited effect on containing the expansion of differences of strengths between agents as established within the first games of matches in which the norm is generally disregarded at the outset (i.e., where

normConcernStrength is less than feedSetpoint). This is reflected in the marked drop in the means of the deviation of strength between normConcernStrength values smaller than feedSetpoint on one hand and the case where normConcernStrength equals feedSetpoint on the other hand for the linear sampling of normConcernStrength used in the evaluation (Figure 7). Analogously, the slope of the plots in Figure 10 flatten only minimally when norm obeyance kicks in.

Table 6: Results of experiment 2

normConcernStrength	**Agg**	st. dev.	**Str**	st. dev.	**Dev**	st. dev.
0	9308	702	4020	90	1722	107
1000	9281	690	4016	83	1720	109
2000	9199	724	4059	89	1680	122
3000	8949	811	4137	100	1650	145
4000	8431	697	4290	89	1571	133
5000	6902	740	4624	80	1531	198
6000	5070	431	5018	37	1467	222
7000	3289	244	5408	44	1413	318
8000	1482	54	5955	19	418	31

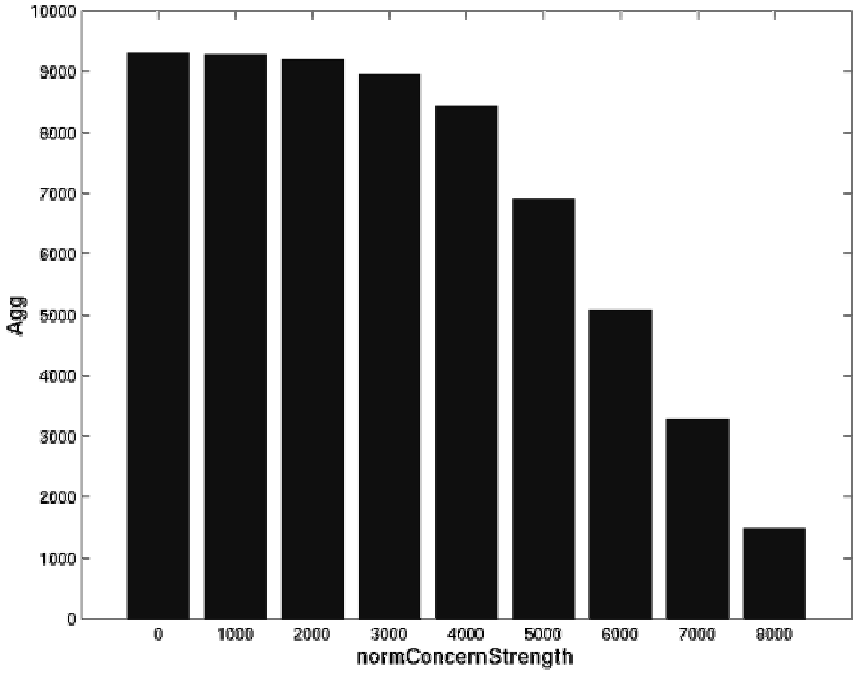


Figure 5: The means of the number of aggressions (**Agg**) for different normConcernStrength values

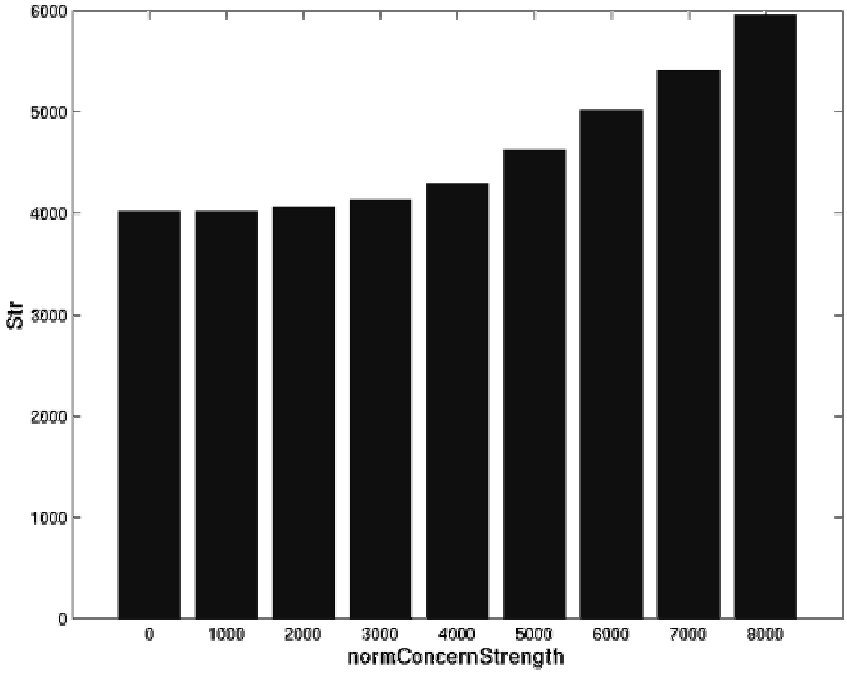


Figure 6: The means of the average strength (**Str**) for different `normConcernStrength` values

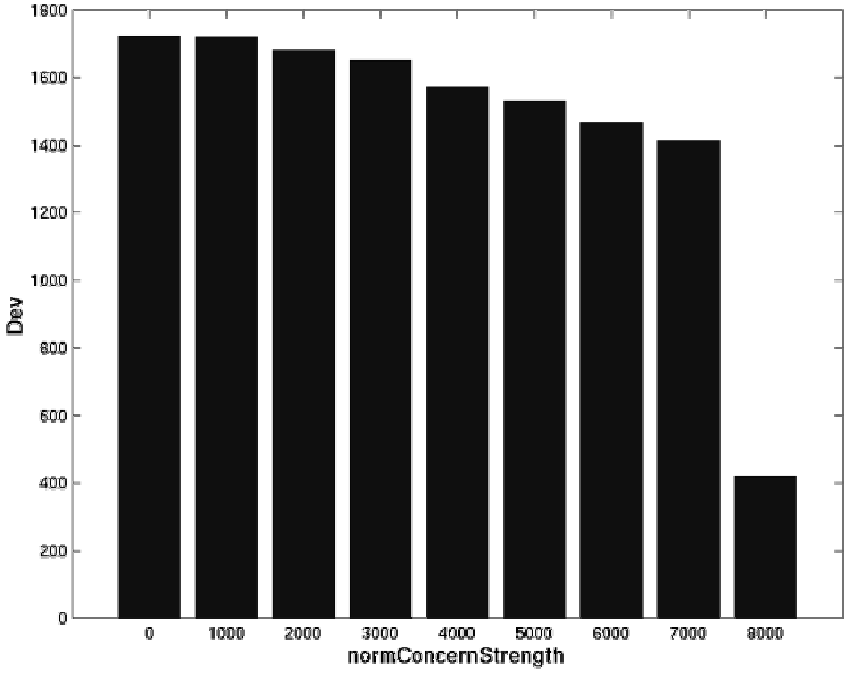


Figure 7: The means of the deviation of strength (**Dev**) for different `normConcernStrength` values

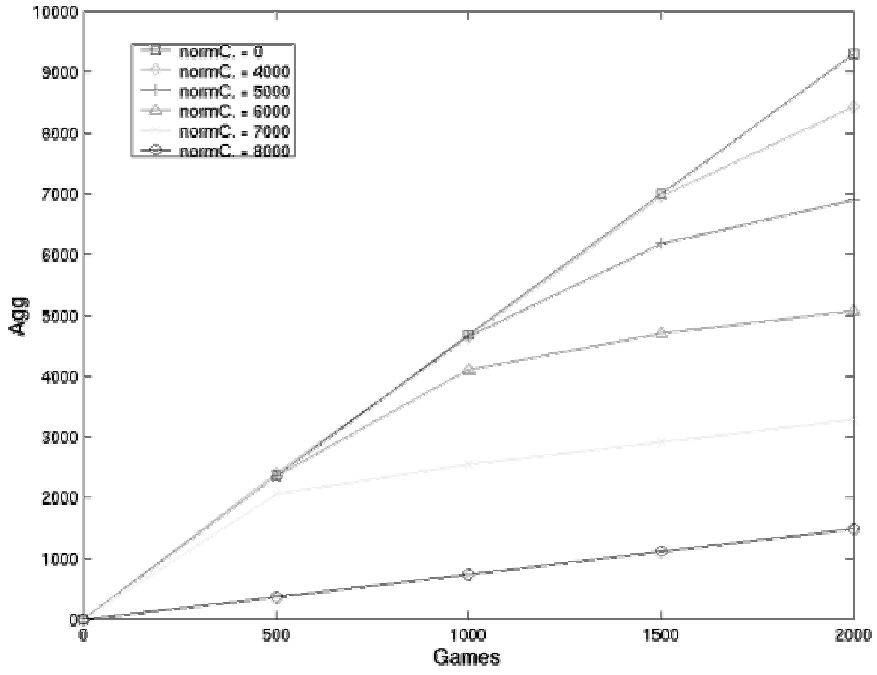


Figure 8: The means of the number of aggressions (**Agg**) after 500, 1000, 1500, and 2000 games for different `normConcernStrength` values

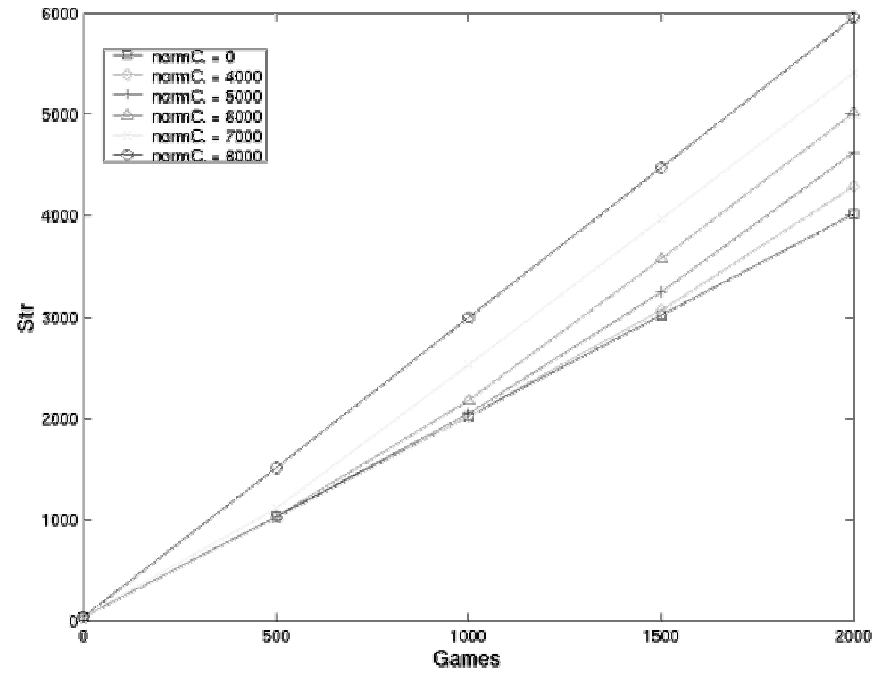


Figure 9: The means of the average strength (**Str**) after 500, 1000, 1500, and 2000 games for different `normConcernStrength` values

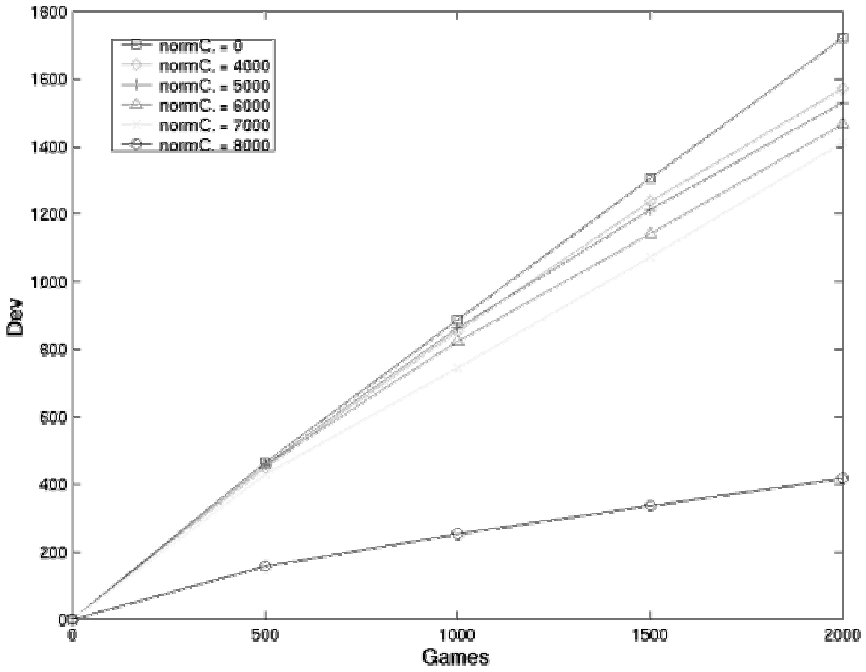


Figure 10: The means of the deviation of strength (**Dev**) after 500, 1000, 1500, and 2000 games for different `normConcernStrength` values

8 Conclusion

This paper is a first step towards including emotions in the computational study of social norms. In humans, emotions are crucial for the efficacy of norms. In our opinion, computational research so far has not paid adequate attention to this aspect. A popular approach towards incorporating norms into the behaviour of agents is the "cognitive" approach (Conte and Castelfranchi 1995a),which is based on the BDI theory. For example, Castelfranchi et al. (2000) propose an agent architecture for "deliberative normative agents" that is based on representations of beliefs, goals, and intentions and enables the agent to "intelligently" violate norms. We propose to include emotions in agent architecture models in order to more adequately capture the influence of norms on behaviour. To this end, we have incorporated emotions into a BDI architecture and have modelled how emotions influence the decision whether to obey or violate norms. In this sense, our proposal can be seen to extend the hitherto declarative cognitive approach by covering some procedural aspects of the emotional system as described in current functional models from psychology.

At the same time, there is another important motivation for us to move away from logic-based models of human behaviour: the theory that has typically been used to incorporate norms into agent architectures is deontic logic (e.g., Dignum 1999). The theoretical rigour offered by logic is certainly appealing. However, if the goal of computational research on norms is to model *human* behaviour, logic does not provide an adequate foundation. Deontic logic must not be equated with human deontic reasoning, which has attracted much interest of psychologists. Staller, Sloman and Ben-Zeev (2000) review theories of deontic reasoning and propose how deontic reasoning may in turn be based on the representation of norm violating instances.

The model described in the present paper is part of TABASCO, a research programme aimed at the implementation and evaluation of tractable architectures for situated cognisers. Therein, emotions and norms are seen to be key elements interconnecting the micro and macro levels of multi-agent instantiations. Building upon the promising results of this evaluation of an appropriately simplified instance of $TABASCO_{JAM}$, there now are numerous entries on our to-do list about how to improve the current implementation. The appraisal component is to be extended towards the coverage of more complex (in particular, also temporal) perceptions, the extension of coverage of different appraisal criteria, and the consideration of specific coping potentials. Similarly, impulse generation is to be improved, in particular with respect to the determination of intensity of the impulse and the generation and management of multiple concurring action tendencies. Action tendencies themselves are in turn to be displayed automatically (cf. Frijda's account of their relationship to facial expressions), providing an important additional channel of communication for inter-agent coordination. Mechanisms for internal regulation and generation of overt as well as cognitive action are further important topics for research. Last but not least, we cannot omit the particular challenge of having to define increasingly complex scenarios that form adequate environments for our situated agents by actually asking for these extended capabilities, while lending themselves to rigorous evaluation.

Looking beyond the narrow limitations of the initial application study discussed in this paper, there are various aspects of the model presented that hold promise of a broader relevance for the domain of computational systems as societies (in addition to the critique of the application of deontic logic given above). Assigning “emotions” their proper role in the realization of a society enables a systematic modularisation of its control system by the separation of (not outcome-oriented) norms and emotions that provide flexible connection of norms to different actual *stories* (see section 1.1). This resulting combination can candidate as a means to complement bounded rationality and overcome some of its most prominent problems, including e.g. the tragedy of the commons (Turner, 1993) or the prisoner’s dilemma (section 1.1.2), without having to resort to overly restrictive and static measures. Emotions, as discussed, encompass mandatory involuntary processes that apply to *both* sending and receiving ends. The information therein processed encodes the interpretation of the momentary status quo, which is interwoven with the individual’s views of what future developments are deemed possible. The abstraction captured in the notion of an “action tendency” (section 2.2) in appraisal theory allows for communication of meaningful content while at the same time allowing for ample variety in individual handling of situations. This also facilitates the construction and application of useful observation-based models of others. While the computational requirements, especially at the individual level, may seem to be all but downright preposterous at first sight, we point out that the choice of a full-blown BDI architecture as a basis for $TABASCO_{JAM}$ was driven by utilitarian considerations as mentioned at the end of section 4. In the realization of (more) specialized, deployment-ready agent systems, many of the abstractions referred to by appraisal theory can be readily mapped to well-known standard components, or combinations or components thereof. Even so, it is certainly true that the present considerations presuppose scenarios with individuals of a certain complexity and are not to be taken to be of universal applicability. For scenarios meeting the due caveats, our ongoing investigations aim to ascertain the feasibility of providing a “softer” model of communication and control that more easily accommodates a varied populations.

References

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

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

CASTELFRANCHI C, Conte R and Paolucci M (1998) Normative reputation and the costs of compliance. *Journal of Artificial Societies and Social Simulation*, 1(3). <http://www.soc.surrey.ac.uk/JASSS/1/3/4.html>

CASTELFRANCHI C, Dignum F, Jonker C M and Treur J (2000) Deliberative normative agents: Principles and architecture. In Jennings N R and Lesperance Y, *Intelligent Agents VI. Agent Theories, Architectures, and Languages, Proc. 6th International Workshop, (ATAL'99)* Orlando, Florida, USA. Berlin/Heidelberg: Springer-Verlag, LNCS 1757.

CONTE R and Castelfranchi C (1995a) *Cognitive and Social Action*. London: UCL Press.

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

DIGNUM F (1999) Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1). pp. 69-79.

EKMAN P and Friesen W V (1975) *Unmasking the Face*. Englewood Cliffs, NJ: Prentice Hall.

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

ELLIOTT C (1997) Hunting for the Holy Grail with "emotionally intelligent" virtual actors. *ACM Intelligence*, 1(1).

ELSTER J (1996) Rationality and the emotions. *The Economic Journal*, 106(438). pp. 1386-1397.

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

FIRBY (1989) *Adaptive Execution in Complex Dynamic Worlds*, PhD Thesis. New Haven, CT: Yale University.

FORSYTH D R (1994) Norms. In Manstead T and Hewstone M (Eds.), *Blackwell Encyclopedia of Social Psychology*, Oxford, UK: Blackwell.

FRANK R H (1988) *Passions within Reason: The Strategic Role of the Emotions*. New York: Norton.

FRIJDA N H (1986) *The Emotions*. Cambridge, UK: Cambridge University Press.

FRIJDA N H, Ortony A, Sonnemans J, Clore G L (1992) The complexity of intensity: Issues concerning the structure of emotion intensity. In Clark M S (Ed.), *Emotion. Review of Personality and Social Psychology (Vol. 13)*, Newbury Park, CA: Sage. pp.60-89.

FRIJDA N H (1995) Emotions in robots. In Roitblat H L and Meyer J-A (Eds.), *Comparative Approaches to Cognitive Science*, Cambridge, MA: MIT Press. pp. 501-516.

GAT E (1997) On three-layer architectures. In Kortenkamp D, Bonasso R P and Murphy R (Eds.), *Artificial Intelligence and Mobile Robots*, Cambridge, MA: MIT Press.

GEORGEFF M P and Lansky A L (1987) Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, WA. pp. 677-682.

HOCHSCHILD A R (1983) *The Managed Heart: Commercialization of Human Feeling*. Berkeley, CA: University of California Press.

[HORSWILL] I. (1997) Horswill I (1997) Visual Architecture and cognitive architecture. In Hexmoor H.H. (Ed.), *Journal of Experimental and Theoretical Artificial Intelligence*, 9(2/3), 277-292.

HUBER M J (1999a) JAM: A BDI-theoretic mobile agent architecture. In Etzioni et al. (Eds.) *Proceedings of the Third International Conference on Autonomous Agents (Agents'99)*, Seattle, WA. pp. 236-243.

HUBER M J (1999b) *JAM Agents in a Nutshell*. Oceanside, CA: Intelligent Reasoning Systems. < http://members.home.net:80/marcush/IRS/Jam/Jam-man.html>

KELTNER D and Buswell B N (1997) Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin*, 122. pp. 250-270.

KELTNER D and Gross J J (1999) Functional accounts of emotions. *Cognition and Emotion*, 13(5). pp. 467-480.

KELTNER D and Haidt J (1999) Social functions of emotions at four levels of analysis. *Cognition and Emotion*, 13(5). pp. 505-521.

LAZARUS R S (1991) *Emotion and Adaptation*. Oxford, UK: Oxford University Press.

LEE J, Huber M J, Durfee E H and Kenny P G (1994) UM-PRS: An implementation of the Procedural Reasoning System for multirobot applications. In *Proceedings of the Conference on Intelligent Robotics in Field, Factory, Service, and Space (CIRFFSS '94)*, Houston, TX. pp. 842-948.

LEVENSON R W (1999) The intrapersonal functions of emotion. *Cognition and Emotion*, 13(5). pp. 481-504.

MELTON R J, Clore G L and Ortony A (1993) *Cognitive Determinants of Emotional Intensity*. Paper presented at the sixty-fifth annual meeting of the Midwestern Psychological Association, Chicago, IL.

MYERS K L (1997) *User Guide for the Procedural Reasoning System*, Technical Report. Menlo Park, CA: SRI International, Artificial Intelligence Center.

ORTONY A, Clore G L and Collins A (1988) *The Cognitive Structure of Emotions*. Cambridge, UK: Cambridge University Press.

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

PETTA P, Staller A, Trappl R, Mantler S, Szalavari Z, Psik T and Gervautz M (1999) Towards engaging full-body interaction. In Bullinger H-J and Vossen P H (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)*, Stuttgart, Germany: Fraunhofer IRB Verlag. pp. 280-281.

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

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

ROSEMAN I J, Antoniou A A and Jose P E (1996) Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion*, 10(3). pp. 241-278.

RUSSELL S J (1997) Rationality and intelligence. *Artificial Intelligence, Special Issue on Economic Principles of Multi-Agent Systems*, 94(1-2). pp. 57-77.

RUSSELL S J and Norvig P (1995) *Artificial Intelligence - A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.

SAAM N J and Harrer A (1999) 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>

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

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

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

SCHERER K R (1999) Appraisal theory. In Dalgleish T and Power M (Eds.), *Handbook of Cognition and Emotion*, Chichester: Wiley. pp. 637-663.

SEIF EL-NASR M, Yen J and Ioerger T R (2000) FLAME - A fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, 3(3).

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

SHOHAM Y and Tennenholtz M (1992b) 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*, San Mateo: Kaufman. pp. 225-231.

SIMON H A (1967) Motivational and emotional controls of cognition. *Psychological Review*, 74. pp. 29-39.

SMITH C A and Ellsworth P C (1985) Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48. pp. 813-838.

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

SONNEMANS J and Frijda N H (1994) The structure of subjective emotional intensity. *Cognition and Emotion*, 8(4). pp. 329-350.

SONNEMANS J and Frijda N H (1995) The determinants of subjective emotional intensity. *Cognition and Emotion*, 9(5). pp. 483-506.

SOUSA R de (1987) *The Rationality of Emotion*. Cambridge, MA: MIT Press.

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

STALLER A, Sloman S A and Ben-Zeev T (2000) Perspective effects in nondeontic versions of the Wason selection task. *Memory & Cognition*, 28(3). pp. 396-405.

TANGNEY J P (1999) The self-conscious emotions: shame, guilt, embarrassment and pride. In Dalgleish T and Power M (Eds.), *Handbook of Cognition and Emotion*, Chichester: Wiley. pp. 541-568.

THOITS P A (1990) Emotional deviance: Research agendas. In Kemper T D (Ed.), *Research Agendas in the Sociology of Emotions*, Albany, NY: State University of New York Press. pp. 180-203.

TRIVERS R L (1971) The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46. pp. 35-57.

[TURNER] R.M. (1993) The Tragedy of the Commons and Distributed AI Systems. In *Proceedings of the 12th International Distributed Artificial Intelligence Workshop*, Hidden Valley, PA, also available as UNH CS Technical Report 93-01.

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

ZAJONC R B (1980) Feeling and thinking: Preferences need no inferences. *American Psychologist*, 2. pp. 151-176.

Bibliographic Information

Alexander Staller is a PhD student of computer science at the Dept. of Medical Cybernetics and Artificial Intelligence of the University of Vienna and a member of the "Intelligent Software Agents and New Media" group at the Austrian Research Institute for Artificial Intelligence. He is writing his PhD thesis under the supervision of Prof. Robert Trappl and focuses on modelling emotions among agents situated in social virtual environments. Further fields of interests include software agents in general and human reasoning.

Paolo Petta is the head of the "Intelligent Software Agents and New Media" group at the Austrian Research Institute for Artificial Intelligence and lecturer at the Dept. of Medical Cybernetics and Artificial Intelligence of the University of Vienna.

Address of both authors:
Austrian Research Institute for Artificial Intelligence
Schottengasse 3
A-1010 Vienna
Austria (EU)
email: {alexs, paolo}@ai.univie.ac.at
URL: http://www.ai.univie.ac.at/oefai/agents

Acknowledgements

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Education, Science and Culture.

This research is being carried out under project GZ 61.096/4-V/B/99 of the Austrian Federal Ministery of Science, Transport and the Arts.