Extending the Computational Study of Social Norms with a Systematic Model of Emotions

Ana L. C. Bazzan* and Diana F. Adamatti*
Rafael H. Bordini

Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15064, 91501–970, Porto Alegre, Brazil
{bazzan,adamatti,bordini}@inf.ufrgs.br

Abstract. It is generally recognized that the use of emotions plays an important role in human interactions, for it leads to more flexible decision–making. In the present work, we extend the idea presented in a paper by Castelfranchi, Conte, and Paolucci, by employing a systematic and detailed model of emotion generation. A scenario is described in which agents that have various types of emotions make decisions regarding compliance with a norm. We compare our results with the ones achieved in previous simulations and we show that the use of emotions leads to a selective behavior which increases agent performance, considering that different types of emotions cause agents to have different acting priorities.

Keywords

Social norms, Emotions and personality, Multiagent–based simulation

1 Introduction

There are several arguments suggesting that emotion affects decision–making (see for instance [6] for a discussion on this issue). It is generally recognized that the benefits of humans having emotions encompass more flexible decision–making, as well as creativity. However, little work has focused on the investigation of interactions among social agents whose actions are somehow influenced by their current emotional setting.

Our overall goal is to create a framework to allow users to define the characteristics of a given interaction, the emotions agents can display, and how these affect their actions and interactions. In a previous paper [2], we have presented a prototype of such a framework using the Iterated Prisoner's Dilemma (IPD) scenario as a metaphor for interactions among agents. The present paper describes the use of a systematic model for generation of emotions applied to the scenario proposed in [4], in order to extend the study of the functions of social

* Author partially supported by CNPq

norms, such as the control of aggression among agents in a world where one's action influences the achievement of others' goals.

A review on the ideas motivating our work, namely, the specific scenario proposed by Castelfranchi *et al.* ([4],[5]) is presented in Section 2. The use of emotions in computing is given in Section 3. The proposed framework, its use in that scenario, and the results obtained by the modeling of agents with emotions are presented in Sections 4, 5, and 6, respectively. Section 7 concludes the paper and mentions future directions of the work.

2 The Scenario and Previous Results

The scenario proposed by Conte and Castelfranchi in [4] aimed at studying the effects of normative and non–normative strategies in the control of aggression among agents, and to explore the effects of the interaction between populations following different criteria for aggression control. The world as devised by the authors is a 10×10 square grid with randomly scattered food, in which agents can move in four directions only: up, down, left, and right. Various experiments were carried out (100 repetitions of a match consisting of 2000 time steps), in which the characteristics of agents were defined in different ways, as explained below. In all experiments, agents and food items are assigned locations at random. A cell cannot contain more than one object at a time, except when an agent is eating. At the beginning of each turn, every agent selects an action from its agenda according to the utility each will bring.

Eating is the most convenient choice for an agent. It begins at a given turn and may end two turns later if it is not interrupted by aggression. The eater's strength changes only when eating has been completed, that is, the eater's strength changes in a discrete way. When a food item has been consumed, it is immediately restored at a randomly chosen location in the grid. The second best choice of an agent is to move to a unoccupied grid position in which food has been seen. Agents can see food only within its territory, which consists of the four cells to which an agent can move in one step from its current location. The next choice is to move to a position where food has been smelt (if the agent does not see a food item, it can smell it within its extended neighborhood which consists of two steps in each direction from the agent's current location). Aggression is the next option available to an agent. If no food is available (either by seeing or smelling), an agent may attack an eating neighbor.

The outcome of an attack is determined by the agents' respective strengths (the stronger agent always wins). When the competitors are equally strong, the defender is the winner. The cost of aggression is equal to the cost of being attacked. Agents may be attacked by more than one agent at a time, in which case the victim's cost is multiplied by the number of aggressors. However, in this case only the strongest attacker earns the food item, while the others get nothing. Finally, the two last choices available to an agent are to move randomly (if no food is seen, smelt, and no attack is possible), and to pause (if even a random move is not possible).

Each match of 2000 time steps includes 50 agents and 25 food items with a nutritive value of 20 units each. Initially, agents' strength are set to 40 units. During the match, agents have to pay the costs of their actions: 0 for pausing, 1 for moving to an adjacent cell, 4 for attacking or being attacked.

The main objective of the work on this scenario was the comparison of number of attacks and strength of agents when they follow social norms and when they act according to utilitarian rules. Three types of agents were proposed:

blind (B) whose aggression is constrained only by personal utility, with no reference to the eaters' strength. Blind agents attack eaters each time the cost of alternative action (as explained above) is higher. In other words, they are not aware of the eaters' strengths, nor of their own.

strategic (S) whose aggression is constrained by strategic reasoning. Strategic agents will only attack those eaters whose strength is not higher than their own. An eater's strength is perceptible one step away from the agent's current location.

normative (N) are the ones which follow a norm of precedence regarding agents that find food, thus becoming their owners. Each time agents or food are randomly allocated on the grid, the latter are assigned to the former when they happen to fall into the agents' territories. Owned food items are flagged and every player knows to whom they belong. Normative agents cannot attack possessors eating their own food.

From the results obtained in [4] and [5], normative strategies were found to reduce aggression, and also to afford the highest average strength and the lowest polarization of strength among the agents when compared to non–normative strategies.

Later on, the same scenario was used by Staller and Petta in [12], in order to investigate the interrelation between social norms and emotions. To this end, they adopted the same scenario of the study in [4], except for the simple action selection algorithm of the agents. In order to study the micro–level processes (as Staller and Petta put) underlying the interrelation between social norms and emotions, social simulations with more complex agents whose architecture includes emotion were conducted.

They conclude that emotions are crucial for the efficacy of norms, and that computational research has not yet paid adequate attention to this aspect, as we also pointed out in [3]. However, their claim is based on the idea of appraisal of concern–relevance only. The authors do not exactly use a model of emotions. Rather, they use some *ad hoc* variables measuring the intensity of states the agent is concerned with. For instance, depending on the average strength, the aggression level is modified.

In summary, in their approach, the act of obeying norms or not comes out because the intensities of aggression and strength were modified, not as a consequence of the type of emotions the agents have. This is the main motivation for our present work: we think their work can be improved by the use of a cognitive structure of emotions such as the one proposed in [10] and previously used by us

in [2]. This may not only yield similar qualitative results, but also do so based on more sound grounds.

3 How Emotions Influence Decision and Behavior

The research on human emotions has a long tradition, both on a cognitive as well as on a physiological basis. See for instance the work in [8, 9] for the latter. However, we focus our work on the former, especially on the synergy between research in this field and decision–making, which, in turn, is relevant to many areas of artificial intelligence. In fact, a trend in the direction of agents displaying heterogeneous behaviors is reported in the literature (in quite distinct scenarios).

We do not attempt here to define what emotions are. As Picard [11] puts it, researchers in the area do not even agree on a definition. Rather, we concentrate on the cognitive and behavioral aspects of emotions as a computationally tractable model. This brings us to the need of stating the eliciting conditions for a particular emotion to arise, as well as the actions carried out as a consequence of it.

For our purposes, we find the so–called OCC theory by Ortony, Clore, and Collins [10] the most appropriate one. First, the authors are very concerned with issues dear to the Artificial Intelligence community; for instance, they believe that cooperative problem–solving systems must be able to reason about emotions. This is clearly an important research issue in Multi–Agent Systems as well. Second, it is a very pragmatic theory, based on grouping emotions by their eliciting conditions – events and their consequences, agents and their actions, or objects – which best suits a computational implementation.

The overall structure of the OCC model is based on emotion types or groups, which in their turn are based on how people perceive the world. They assume that there are three major perception aspects in the world: *events*, *agents*, and *objects*. Events are simply people's construal about things that happened (not related to beliefs, nor necessarily with their possible causes). Objects are also a very straightforward level of perception. Finally, agents are both human and nonhuman beings, as well as inanimate objects (as explained next) or abstractions. In short, by focusing on events, objects, and agents, one is interested in their consequences, properties, and actions, respectively. In this model, another central idea is that emotions are valenced reactions; the intensity of the affective reactions determines whether or not they will be experienced as emotions. This points to the importance of framing the variables which determine the intensity of any reaction.

The structure of the OCC model based on types of emotions has three main branches, corresponding to the three ways people react to the world.

The first branch relates to emotions which are arising from aspects of **objects** such as liking, disliking, etc. This constitutes the single class in this branch, namely that called *attraction* which includes emotions such as love and hate.

The second branch relates to emotions which are consequences of **events**. Three classes appear here: *fortunes–of–others* (emotions *happy–for* and *gloating*

or *Schadenfreude*); *prospect–based* (emotions *hope*, which can be either confirmed as *satisfaction* or disconfirmed as *disappointment*, and *fear*, which can be either confirmed as *fears-confirmed* or disconfirmed as *relief*); and *well–being* (emotions *joy* and *distress*).

The third branch is related to consequences of **agents**, namely the *attribution* class, comprising the following emotions: *pride* (person approves of self), *admiration* (person approves of other), *shame* (person disapproves of self), and *reproach* (person disapproves of other).

Finally, an additional class of emotions can be referred to as *compound*, since it focuses on both the action of an agent, and the resulting event and its consequences. This class is called *well–being/attribution compound*. It involves the emotions of gratification, remorse, gratitude, and anger.

Ortony *et al.* [10] recognize that this model is oversimplified, since in reality a person is likely to experience a mixture of emotions, especially when considering a situation from different perspectives at different moments. However, this co–occurrence would probably render the model computationally unfeasible. We believe that the model does have merits when one's goal is to conduct experiments on the effects of focusing on various aspects of an emotion–induced situation (as we do here), rather than to attempt at analyzing exactly what combinations or sequences of emotions could occur in given situations.

As for the intensity of emotions, which is important if one wants to implement a computational model, possibly relating certain variables to thresholds, Ortony *et al.* [10] distinguish between local and global variables affecting such intensity. Global variables affect all the types of emotions they have identified, and include: sense of reality (how much one believes the situation), proximity (how close one feels the situation), unexpectedness (how surprised one is by the situation), and arousal (how much one is aroused prior to the situation). On the other hand, local variables affect only specific groups of emotions. For example, the event–based emotions are affected by the desirability variable.

Some papers report previous usage of the OCC model. Elliott [7] has built the Affective Reasoner to map situations and agent state variables into a set of specific emotions, producing behaviors corresponding to them. Bates [1] has worked on micro–worlds that include moderately competent, emotional agents for the Oz Project.

4 The Proposed Framework

Our overall goal is to create a framework that allows users to define the characteristics of given interactions, the emotions agents can display, and how these affect their actions (hence those interactions). Such a framework is intended to be very general. That is, the user specifies the purpose of the simulation; the scenario for the interactions (which rules or norms agents follow when they meet); the environment (e.g., interactions happen among agents which belong to particular groups, agents are not attached to any group and meet randomly, interactions happen with respect to a spatial/geographical configuration); general parame-

ters of the simulation (time frame, size of environment, etc.); the classification of any emotion that does not belong to the original OCC model, or the whole meaning of an emotion if it does not fit the model at all; and parameters related to each agent in the simulation (thresholds, types, etc.).

We base our framework on the OCC model for the reasons already explained. Additionally, this model can be translated into a rule–based system that generates cognitive–related emotions in an agent. We now explain how the rules look like in such a system. The IF part tests either the desirability (of a consequence of an event), or the praiseworthiness (of an action of an agent), or the appealingness (of an object). The THEN part sets the potential for generating an emotional state (e.g., a joyful state).

Let $A(p,o,t)$ be the appealingness of an object that a person p assigns to the object o at time t, $P_h(p,o,t)$ the potential to generate the state of hate, $G(vg_1,\ldots,vg_n)$ a combination of global intensity variables, $I_h(p,o,t)$ the intensity of hate, $T_h(p,t)$ a threshold value, and f_h a function specific to hate. Then, a rule to generate a state of hate looks like:
IF $P_h(p,o,t) > T_h(p,t)$ THEN set $I_h(p,o,t) = P_h(p,o,t) - T_h(p,t)$ ELSE set $I_h(p,o,t) = 0$

This rule is triggered by another one:
IF $A(p,o,t) > 0$ THEN set $P_h(p,o,t) = f_h(A(p,o,t),G)$

Ortony *et al.* [10] omit many of the details of implementation; a difficult issue might be to find appropriate functions for each emotion. It remains to be investigated whether general functions exist or whether they are domain–dependent. While we are studying these and other questions related to the implementation the OCC structure in a general framework, we are testing them on specific scenarios such as the one on the simulation of social norms.

5 Simulation of the Social Norms Scenario using the OCC Model

The framework presented in the previous section was already used by us [2] in the simulation of a classic scenario concerning the Iterated Prisoner's Dilemma (IPD). It was shown that the use of emotions in such scenario increased the rate of cooperation.

We also maintain that the study of social norms is highly significant in the field of Multi–Agent Systems. Social norms form the basis of an approach to agent coordination (a central issues in MAS), facilitating decision–making by autonomous agents and avoiding unnecessary conflicts. The approach to agent coordination to which we refer is the one based on the social and cognitive sciences, rather than game theory. It is directly inspired by the role that social norms play in human societies. We argued in [3] that emotions are important in attaching agents to social norms, and that including emotions in agent architectures may help us come to a better understanding of how autonomous agents form and perpetuate conventions, which are essential for social behavior.

We have restricted the framework to agents displaying a single emotion (the predominant one), which is a consequence of the use of the OCC model. We start by identifying a set of emotions related to the scenario itself. As a first exercise, we have concentrated on typical emotions. Thus, agents may initially display anger (a), joy (j), resentment (r), and pity (i). This way, almost all classes of the OCC model are represented within the set of emotions we use.

Let us now turn to the IF–THEN–ELSE rules derived for this specific scenario. All variables defined in Section 4 retain their meaning here. Beside those, we use below: $D(p,e,t)$ for the desirability that a person p assigns to event e at time t, $W(p,g,t)$ for the praiseworthiness that a person p assigns to agent g at time t, and $L(vl_1,..,vl_n)$ a combination of local intensity variables.

- Rules for joy:
  IF $D(p,e,t) > 0$
  THEN set $P_j(p,e,t) = f_j(D(p,e,t),G,L)$
  function f_j returns value $T_j(p,t)+\epsilon$ (IF agent's strength $> average\_strength$)
  IF $P_j(p,e,t) > T_j(p,t)$
  THEN set $I_j(p,e,t) = P_j(p,e,t) - T_j(p,t)$
  ELSE set $I_j(p,e,t) = 0$
- Rules for resentment:
  IF $D(p,e,t) < 0$
  THEN set $P_r(p,e,t) = f_r(D(p,e,t),G,L)$
  function f_r returns value $T_r(p,t)+\epsilon$ (IF agent's strength $= average\_strength \pm \delta$ AND some agent is eating food which does not belong to it)
  IF $P_r(p,e,t) > T_r(p,t)$
  THEN set $I_r(p,e,t) = P_r(p,e,t) - T_r(p,t)$
  ELSE set $I_r(p,e,t) = 0$
- Rules for pity:
  IF $D(p,e,t) < 0$
  THEN set $P_i(p,e,t) = f_i(D(p,e,t),G)$
  function f_i returns value $T_i(p,t)+\epsilon$ (IF agent's strength $= average\_strength \pm \delta$ AND eater's strength $< average\_strength$)
  IF $P_i(p,e,t) > T_i(p,t)$
  THEN set $I_i(p,e,t) = P_i(p,e,t) - T_i(p,t)$
  ELSE set $I_i(p,e,t) = 0$
- Rules for anger:
  IF ($D(p,e,t) < 0$ AND $W(p,g,t) < 0$)
  THEN set $P_a(p,e,g,t) = f_a(D(p,e,t),W(p,g,t),G,L)$
  function f_a returns value $T_a(p,t) + \epsilon$ (IF agent's suffered aggression $> average\_aggression$ OR agent's strength $< average\_strength$)
  IF $P_a(p,e,g,t) > T_a(p,t)$
  THEN set $I_a(p,e,g,t) = P_a(p,e,g,t) - T_a(p,t)$
  ELSE set $I_a(p,e,g,t) = 0$

We now explain the rules. A joyful agent is defined as one whose strength is above the average (computed over all agents). There are two conditions under which an agent displays resentment: when its strength is close to the average (we

can vary this threshold by changing the parameter δ), or when it perceives some agent eating other agents' food. The definition of a pitiful agent is as follows: its strength is close to the average, and it sees other agent(s) whose strength is below the average. Finally, angry agents are those whose suffered aggression is higher than the average or whose strength is lower than the average strength.

Once fired, emotions have the following effects: joyful agents do not eat or attack (they only move at random); agents feeling resentment attack any agent eating others' food (regardless of strength); pitiful agents do not attack agents eating others' food if their strength is below the average; and angry agents never obey the norm: they attack any eating agent they perceive. Emotions are allowed to fire only after 200 steps of simulation, during which agents behave normatively.

6 Results and Comparison

Several comparisons can be made between the previous simulations of this scenario and the one we proposed here. Initially, we have replicated the experiments as proposed in [5]. We exclude the simulation of the strategic agents because they are of little significance for the comparison of different implementations of normative agents. Discrepancies can be explained by different interpretations of the scenario (an issue also reported in [12]).

Table 1 shows the results of our simulations. The first two lines show the replication of the results reported in [5] for blind and normative agents. The last line contains the results for the simulation with emotions. "Str." is the average strength over the 50 agents. "Dev." is the standard deviation regarding this average, and "Aggr." is the sum of aggressions suffered by the 50 agents. Each of these quantities is associated with a standard deviation ("dev.") computed over 100 repetitions of the simulation (for 2,000 time steps).

Table 1. The results of our simulations.

Type	Str.	dev.	Dev.	dev.	Aggr.	dev.
blind	4135	191	3669	153	9299	449
normative	6757	29	132	20	2638	86
emotions	4307	223	173	72	2891	141

Next, an evaluation of the performance of our agents in the simulation can be made by comparing our results with those in [12]. Staller and Petta have reported a level of strength between 4624 and 5408 and a level of aggression between 3289 and 6902. In our experiments, the aggression decreased to 2891 (Table 1) since joyful agents never attack, and pitiful ones do this at a small rate. On the other hand, due to the design of the joyful agents (once they are satiated they neither attack nor eat), the strength is relatively low (4307). In fact, since in the beginning of the simulations there are many joyful agents (time steps 200 to 1200), the food items they "reject" account for the difference in strength.

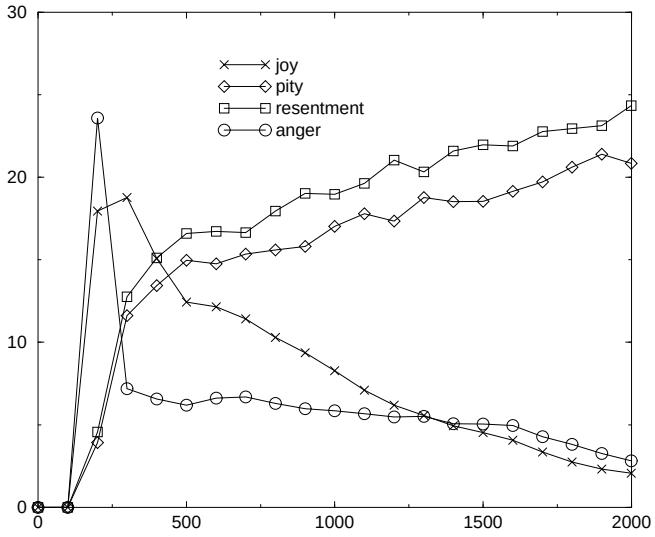

Fig. 1. Number of agents displaying each type of emotion within time.

Finally, the change in number of agents displaying each type of emotion within time is shown in Figure 1. It can be seen that in the beginning of the simulation (up to time step 200) there are only neutral agents because no emotion rule is allowed to fire. After this point, the numbers of angry and joyful agents are high, indicating a polarization of the group of agents. As time passes, resentment and pity increase as a reaction to this situation. This should be better understood in future extensions of the simulation, such as an accounting of the strength and aggression by type of agent.

7 Conclusion and Future Directions

The importance of emotions for human beings is that they yield more flexible decision–making. Staller and Petta have reached a similar conclusion (in [12], regarding the scenarios described in [4] and [5]). Their motivation was that agents do not spend all their time searching for food and attacking other agents. Following this argument, they were able to show that the performance of normative agents improves. We have shown that a similar study may be better conducted based on a formal theory of the cognitive structure of emotions such as that reported in [12]. Therefore, our aim in the present paper has been to carry this out and compare the results.

The present work contributes to the construction of a framework for simulating agents with emotions, by employing a scenario we regard as very important since it deals with social norms for agents. In order to implement that framework, we have first concentrated on finding a computationally tractable model that could account for the cognitive and behavioral aspects of emotions. For our

purposes, we find the so–called OCC model [10] the most appropriate one, especially due to its pragmatical aspects. To prove that the OCC model is suitable for the social norm scenario, we have discussed its structure, as well as some issues which were missed in [10]. This paper also contributes to a deeper understanding of the OCC model regarding implementation details.

Our future plans include the construction of a general framework for simulating user–defined interactions. In order to achieve this, we are defining a series of primitives that can be combined by the users in constructing their own environments. These primitives comprise: the specification of the interactions, which actions to perform, when, and by whom, the wealth of agents, the types of emotions available (both those included in the OCC model and others), among other things. While no such domain–independent rules are available, the users are asked to construct the rules themselves, by entering the parameters for the primitives we have made available.

References

1. J. Bates. The role of emotion in believable agents. In *Communications of the ACM, Special Issue on Agents*, July,1994.
2. A. L. Bazzan and R. H. Bordini. A framework for the simulation of agents with emotions: Report on experiments with the iterated prisoner's dilemma. In J. P. Müller, E. Andre, S. Sen, and C. Frasson, editors, *Proceedings of The Fifth International Conference on Autonomous Agents (Agents 2001), 28 May – 1 June, Montreal, Canada*, pages 292–299. ACM Press, 2001.
3. A. L. Bazzan, R. H. Bordini, and J. A. Campbell. Moral sentiments in multi-agent systems. In J. P. Müller, M. P. Singh, and A. S. Rao, editors, *Intelligent Agents V, Proceedings of ATAL-98*, number 1555 in LNAI, pages 113–131, Heidelberg, 1999. Springer-Verlag.
4. C. Castelfranchi and R. Conte. Understanding the effects of norms in social groups through simulation. In G. N. Gilbert and R. Conte, editors, *Artificial Societies: the computer simulation of social life*, pages 252–267. UCL Press, London, 1995.
5. C. Castelfranchi, R. Conte, and M. Paolucci. Normative reputation and the costs of compliance. *Journal of Artificial Societies and Social Simulation*, 1(3), 1998. <http://www.soc.surrey.ac.uk/JASSS/1/3/3.html>.
6. A. Damasio. *Descartes's Error*. Avon, New York, 1994.
7. C. Elliott. Multi-media communication with emotion-driven believable agents. In *AAAI Spring Symposium on Believable Agents*, Stanford University in Palo Alto, California, March 21-23, 1994.
8. W. James. What is an emotion? *Mind*, 9:188–205, 1884.
9. W. James. *The Principles of Psychology*. Holt, New York, 1890.
10. A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, UK, 1988.
11. R. W. Picard. *Affective Computing*. The MIT Press, Cambridge, MA, 1997.
12. A. Staller and P. Petta. Introducing emotions into the computational study of social norms: A first evaluation. *Journal of Artificial Societies and Social Simulation*, 4(1), 2001. <http://www.soc.surrey.ac.uk/JASSS/4/1/2.html>.