Will Artificial Emotional Agents Show Altruistic Punishment In The Public Goods Game?

Dirk M. Reichardt

Berufsakademie Stuttgart - University of Cooperative Education
D-70180 Stuttgart, Germany
reichardt@ba-stuttgart.de

Abstract: In a public goods game with punishment option free riders will likely be forced by the others to follow the rules. The punishing action is not free of costs. In experiments with human players, punishment also takes place if the players are not scheduled to meet again in a game constellation. This makes the punishment action irrational and it is given the attribute 'altruistic'. Emotions are identified as reasons for the phenomenon of altruistic punishment which contradicts the theory of the homo oeconomicus. The intention of the research work presented here is to model the behaviour of players so that their emotion and personality are bases for decisions and can be recognised as such. In this paper a model of an emotional agent is derived from experiments. The emotion generation is based upon the model of Ortony, Clore and Collins and brings this approach together with a basic model of personality. A first implementation shows plausible behaviour in a simulation environment.

Introduction

In this approach of modelling emotional behaviour, the public goods game (PGG) serves as a scenario in which humans show emotion driven decisions. An agent model which eventually shows similar behaviour can therefore claim to be emotional. The scenario is very limited and only a few decisions are required. Nevertheless, emotional behaviour can be shown and nicely imitated by an artificial player agent.

In the public goods game the punishment option leads to higher average payments (see [1]). It has an educational effect, but it also costs the punisher a certain amount of money. The rationality of this action is based on the assumption that the game is iterated and the behavior of the free rider is changed by the punishment. Fehr and Gächter [1] describe a scenario in which they repeat the game with the condition that the group composition changes from period to period. This makes the punishment decision altruistic, since (only) others may benefit from it. The presence of altruistic punishers is therefore a reason even for selfish players to raise their contributions to the common project. Fehr and Gächter showed a significant correlation between the emotion *anger* and the willingness to punish. How could an artificial emotional agent look like which behaves humanlike in the described scenario? This paper describes an approach for modeling an emotional player which shows plausible reactions.

Outline of the agent architecture

First the typical behaviours of human players are identified. These behaviours have rational and emotional background and serve as role model and orientation for simulated agents. The architecture of the emotional agent includes the generation of emotions and the influence of emotions on decisions. An often quoted cognitive theory of emotion is given by Ortony, Clore and Collins (OCC) [2]. In this approach, the OCC model is adapted to the PGG. In addition, the resulting model is enhanced by a personality model and the representation of mood.

The simulation scenario

The simulation scenario is closely related to the one used by Fehr and Gächter [1]. The core configuration of the public goods game is chosen as follows: each player gets the amount of 20 Euro. The public project is bearing 60 % interest. A player can invest up to 10 Euro to punish any of the other players. Each invested Euro results in a punishment of 3 Euro. The game is repeated for 12 rounds in the same constellation before groups are recombined.

Human player behaviour

In a first step experiments and interviews with (non-experienced) human players are carried out. Abstraction leads to the following typical behaviour patterns:

- Start with a contribution of 15-20, no punishing after the first round, later if only one single candidate tries a free ride (but only by low amounts)
- Start with a contribution of 0-5, do not punish, raise contribution in the next round if punished and less than average income is the consequence
- Start with a contribution of 10, raise contribution if punished, reduce the contribution if unpunished, do not punish

The experiment reveals that the expected punishment for a free rider averages 3.3 Euro. This makes the situation worse for the free rider compared to a complying agent. Even though the players claim to follow the mentioned strategies, some of the observed actions cannot be explained by the scheme:

- A player who usually does not punish or only punishes by low amounts, assigns a rather high punishment amount to another player because this player pays nothing after a few rounds of complete compliance.
- A player who usually always pays, skips one round.

It is assumed that these exceptions are result of an emotional disturbance of the subjectively rational strategy the agent chose. Our model of an emotional agent should be capable of generating similar behaviour. As a first step in the simulation, the emotional agent performance is compared to the performance of agents following the above strategies.

Emotion generation

The architecture of the player agent integrates the emotion generation and the action decisions. A further requirement is the interface to an emotion expression module which is needed for future experiments in which the reaction of humans to emotional agents will be tested.

The OCC model has become the standard model for emotion synthesis. It specifies 22 emotion categories and groups them in three dimensions: *goal relevant events*, *actions of agents* and *aspects of objects*. In the presented model, first the intensities of the predefined eliciting events or actions are derived from the scenario. Each of them is associated to an emotion category. The personality model provides non-linear transfer functions which generate emotion intensity from the current mood and the elicitor intensity.

Two *events* take place in the game: the publication of the payments and of the punishments. Depending on the *goal* of a player (*maximise* own income or get *higher* income than others) the 'well-being'-emotions *joy* and *distress* are generated. The elicitor's intensity is given by a distance function comparing result and goal. The agent model records previous games and derives the *likelihood* of future game situations. Two decisions are made: the payment and the punishment. For each option, a likelihood of the future situations is computed as a (rational) basis for decision. Attached to the likelihood, the emotions *fear* and *hope* are generated. Two factors are considered by a generation-function for those emotions: likelihood and desirability, the latter again based upon the goals of the agent. These emotions are generated before the steps payment and punishment and are (kind of) transformed into the emotions satisfaction (goal achieved), fears-confirmed ("I knew they would punish me for this!"), disappointment ("They should have paid more!"), relief ("I really got through with my trick without getting punished!"). The current model disregards 'fortunes-of-others'-emotions.

The *actions* of the agents are also elicitors for emotion categories computed by the same principle as events. The main factor for the elicitor intensity is the praiseworthiness of the action, which depends on the (moral) standards of the agent. Paying the full amount may be considered praiseworthy since it is beneficial for all. For example: paying nothing whereas all others pay 20 leads to joy but also to shame. Just the degree of shame depends on the personality and therefore the transfer function from the eliciting intensity to the emotion is provided by the personality model. An important point is raised in the example: compound emotions. Especially interesting for this game is the combination of reproach and distress which reveals anger and which in turn can cause a punishment. On the other hand shame (see above) and distress (high punishment) leads to remorse and as a consequence to higher compliance in the next round.

The *attraction* of another player plays a role in the decisions. Here, the elicitor is based upon the benefit/cost caused by the agent in earlier games. Again, the personality model provides a mapping from the elicitor intensity to the emotion intensity including the current mood.

The basics of a personality model

A *mood* can be regarded as a long term emotional state influencing the behaviour. In our case it is modelled on the good-bad scale only. It gets better by positive (e.g. joy) and worse by negative (e.g. distress) emotional experience. How does personality affect the actions? It affects the intensity of the emotions – at least the intensity with which they are shown. Using aspects of Eysencks personality theory [3], the agent's personality model is described by the attributes *openness* and *energy*. Energy describes how *fast* the mood changes and how *high* the amplitudes of the emotion intensities are. Openness describes to what extent experiments (like skipped payments) are possible options.

Emotional and rational decisions

Once the emotion is generated, its influence on the player's decisions is needed. In the selected game context emotions are regarded as input to a (flat) rule based system. This system contains two kinds of rules: direct emotional reactions and reactions with 'deliberation'. If a rule of the first type fires, it cuts off the deliberation. As an example for the second type, the punishment of a player which the agent *hates* a lot has to be decided. In case this agent performs an action which is *disapproved* (does not comply again) it is increased depending on the intensity of the hatred.

Conclusion and future work

On the one hand the model described above allows a lot of adjustment by using parameters and functions for the personality. On the other hand its straight-forward realisation does not make it a generally applicable model of emotional behaviour. As a result of the experiments, watching the actions of the player agent may subjectively suggest the presence of emotions already but closer analysis is needed.

Further steps include adding 'fortunes-of-others'- emotions. Current work is done on the visualisation of the generated emotion by an animated face. This includes a further aspect of the emotional model: the *time*. How fast are emotions generated and how long does it take them to fade? The overall intention is to verify the plausibility of the generated emotion by experiments with human players.

References

[1] Fehr, E., Gächter, S., "Altruistic Punishment in Humans", Nature, Volume 415, pp. 137–140., 2002

[2] Ortony, A., Clore, G., Collins, A., "The Cognitive Structure of Emotions", Cambridge University Press, 1988

[3] Eysenck, H.J., "The structure of human personality", 3rd edition, Methuen, London, 1970