To appear in Intelligent Virtual Agents 2009, Sep 14-16, Amsterdam

Expression of Moral Emotions in Cooperating Agents*

Celso M. de Melo[1], Liang Zheng[2] and Jonathan Gratch[1]

[1] Institute for Creative Technologies, University of Southern California,
13274 Fiji Way, Marina Del Rey, CA 90292, USA
demelo@usc.edu, gratch@ict.usc.edu
[2] Information Sciences Institute, University of Southern California,
4676 Admiralty Way, Suite 1001, CA 90292
liang.zheng@usc.edu

Abstract. Moral emotions have been argued to play a central role in the emergence of cooperation in human-human interactions. This work describes an experiment which tests whether this insight carries to virtual human-human interactions. In particular, the paper describes a repeated-measures experiment where subjects play the iterated prisoner's dilemma with two versions of the virtual human: (a) neutral, which is the control condition; (b) moral, which is identical to the control condition except that the virtual human expresses gratitude, distress, remorse, reproach and anger through the face according to the action history of the game. Our results indicate that subjects cooperate more with the virtual human in the moral condition and that they perceive it to be more human-like. We discuss the relevance these results have for building agents which are successful in cooperating with humans.

Keywords: Moral Emotions, Virtual Humans, Expression of Emotions, Cooperation, Prisoner's Dilemma

1 Introduction

The expression of moral emotions has been argued to influence the emergence of cooperation in human-human interactions [1]. Moral emotions are associated with the interests or welfare of either society as a whole or people other than the self [2]. They show disapproval of another's actions (reproach and anger), regret for one's own actions (shame and remorse) and praise for someone else's action (admiration and gratitude). To understand the effect of moral emotions on emergence of cooperation, consider first the decision model of self-interested agents. In this model agents cooperate only if that improves their own condition, without regard to the other agents' welfare. Now, even though attractive in its simplicity, researchers were quick to notice that people more often than not consider the welfare of others and cooperate [3]. But, cooperation isn't blind and will tend to emerge in situations where

* This work was sponsored by the U.S. Army Research, Development, and Engineering Command. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

participants are willing to cooperate as opposed to trying to take advantage of each other. Therefore, being able to identify when someone is willing to cooperate is key to the emergence of mutual cooperation. Moral emotions take on an important role in this identification process as their display constitutes a cue that someone might be considerate of the welfare of others and is willing to cooperate [1,4].

This paper explores whether the expression of moral emotions can also have an effect on the interaction between people and embodied agents. Embodied agents, or virtual humans, are a special kind of agents which have bodies and are capable of expressing themselves through gesture, face and voice [5]. There has already been much work in trying to promote cooperation, through trust- and reputation-building models, in computational multi-agent systems [6]. However, agents in these systems are optimized to cooperate with other agents. With embodiment, agents can now promote cooperation-building mechanisms which rely on nonverbal behavior, as in human-human interactions [7]. However, whether these mechanisms carry to human-virtual human interactions is not clear. Evidence exists that people interact with machines in similar ways as with people [8]. Further evidence has also been provided that embodied agents can induce social-emotional effects similar to those in human-human interactions [9]. This work seeks evidence that the expression of moral emotions in virtual humans influences people's willingness to cooperate with it.

The paper describes an experiment where humans are asked to play the iterated prisoner's dilemma game with two virtual humans that follow the same action policy but differ only in that one expresses moral emotions and one does not. Expression of moral emotions consists of expressing gratitude, distress, remorse, reproach and anger through the face according to how the game is unfolding. We expect this manipulation to produce an effect on the subjects' willingness to cooperate. The goals of the experiment are to get insight into: (a) whether the expression of moral emotions has an effect on the emergence of cooperation between virtual humans and humans; (b) the magnitude of such effect, if it does exist; and, (c) the importance of embodiment for the emergence of cooperation between agents and humans.

2 Method

Design. The experiment follows a within-subjects design where each participant plays the iterated prisoner's dilemma with two different virtual humans. The virtual humans differ in their expression of moral emotions: a *neutral virtual human*, (the control condition) expresses no emotions; a *moral virtual human* expresses moral emotions. After playing two rounds of the prisoner's dilemma to familiarize themselves with the game, participants play twenty-five rounds of prisoner's dilemma with each agent. The order of the agents is randomized across participants (i.e., one half of the subjects play the neutral virtual human first whereas the other half play the moral virtual human first).

The Game. The iterated prisoner's dilemma game was chosen because it is regularly used in game theory to understand cooperative behavior between two agents [10]. The (non-iterated) prisoner's dilemma game was described in the experiment as follows:

“You and your partner are arrested by the police. The police have insufficient evidence for a conviction, and, having separated you both into different cells, visit each in turn to offer the same deal. If one testifies for the prosecution against the other and the other remains silent, the betrayer goes free and the silent accomplice receives the full 3-year sentence. If both remain silent, both prisoners are sentenced to only 3 months in jail for a minor charge. If each betrays the other, each receives a 1-year sentence”. The two actions in this game were presented to the subject as ‘Remain SILENT’ or ‘TESTIFY against other’. However, in the rest of the paper, we shall refer to the first as ‘cooperate’ and the second as ‘defect’. According to the self-interested model of agents, the only rational action strategy is to defect as this maximizes the expected utility. In the iterated version of the prisoner’s dilemma, the game is played several times. Importantly, in our experiment the game is played a finite number of times (25) and subjects are aware of this. Furthermore, subjects are told that they will be playing each round with the same partner (i.e., virtual human) and that each will have the opportunity of learning what the other did in the previous round. Again, the self-interested model states that the optimal strategy for the finite iterated prisoner’s dilemma is to defect in every round. Finally, subjects were instructed to play the game as if they were actually experiencing the dilemma and to follow the best strategy they saw fit. In particular, subjects were not told about the strategies predicted by the self-interested model.

The Action Policy. Virtual humans in both conditions play a variant of tit-for-tat. Tit-for-tat is a strategy where a player begins by cooperating and then proceeds to repeat the action the other player did in the previous round. Tit-for-tat is argued to strike the right balance of punishment and reward with respect to the opponent’s previous actions [11]. So, the action policy used in our experiment is as follows: (a) in rounds 1 to 5, the virtual human plays randomly; (b) in rounds 6 to 25, the agent plays tit-for-tat. Importantly, the random sequence of actions in the first five rounds is the same in both conditions and is chosen at the beginning of each trial with a new subject. The rationale for having some randomness in the first rounds was to make it harder for the subjects to guess the virtual humans’ strategy.

Expression of Emotions. The moral virtual human expresses emotions after each round of the game. The emotion which is expressed reflects not only the outcome of the last round but also the outcome of (recent) rounds in the past. The way we map the outcome history of the game into emotions follows the eliciting conditions for moral emotions as described by Haidt [2]. The mapping we propose is not meant to be the correct one but only one which is intuitive and reasonable. The mapping is described by the following ordered rules:

a. If in the present round both players cooperate, gratitude is expressed;
b. If in the present round the subject defects and the virtual human cooperates:
    1. If in the previous two rounds both cooperated, anger is expressed;
    2. If in the previous round both cooperated, reproach is expressed;
    3. If in the previous round the subject defected and the virtual human cooperated, reproach is expressed;
    4. Otherwise, distress (or sadness) is expressed;

c. If in the present round the subject cooperates and the virtual human defects, remorse is expressed;

d. If in the present round both players defect:

    1. If in the previous round the subject cooperated and the virtual human defected, don't express any emotion;
    2. Otherwise, express distress.

Virtual Humans. The virtual human platform we use supports expression of emotions through gesture, face and voice [12]. The focus on this work, however, is on expression of moral emotions through the face. Facial expression relies on a pseudo-muscular model of the face and on simulation of wrinkles and blushing. The facial expressions for the moral virtual human condition are shown in Fig.1. These expressions are elicited after the subject chooses its action and the outcome of the round is shown. The neutral condition uses the neutral face. Aside from facial expression, Perlin noise to the neck and torso and blinking was added to both conditions to keep the virtual human from looking stiff while the subject is choosing its actions. Finally, a different shirt (i.e., texture) is chosen to help distinguish the two virtual humans the subject plays with. Shirts vary according to letter (A or B) and color (blue or yellow). The letter 'A' is always assigned to the first player and the letter 'B' to the second. The shirt colors are assigned randomly to each player. Neutral images of the respective virtual humans, with the respective shirts, are also shown in the debriefing questions to help subjects remember the players.

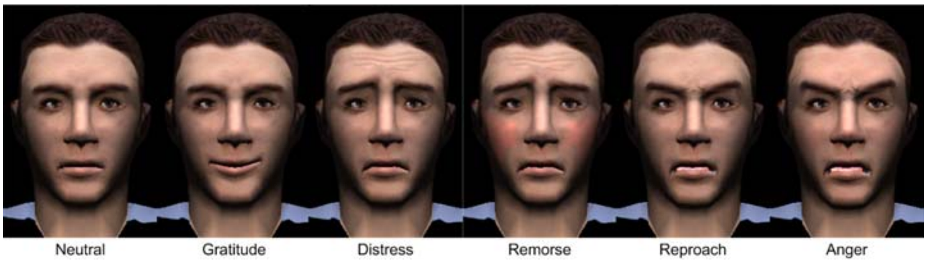
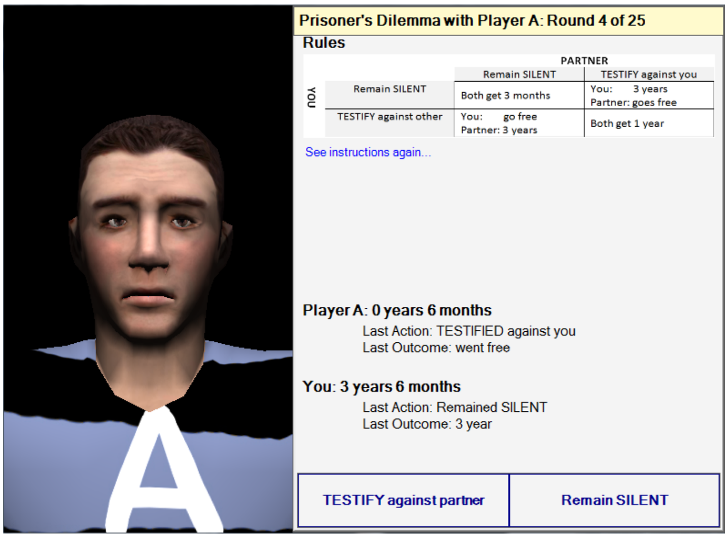


Fig. 1. The facial expressions used for the moral virtual human condition. Usual facial configurations are used for gratitude, distress, remorse, reproach and anger. Typical wrinkle patterns are used for distress, remorse, reproach and anger. Blushing of the cheeks is used for remorse and (light) redness of the face is used in anger.

Survey Software. The survey was implemented in software and structured into four phases: (1) *profile*, where data about the subject is collected (e.g. age, sex and education-level) while assuring anonymity; (2) *tutorial*, where the subject plays a two-round game of the iterated prisoner's dilemma to get comfortable with the game and interface; (3) *game*, where the subject plays the 25-round iterated prisoner's dilemma twice, once for each condition; finally, (4) *debriefing*, where the subject answers the debriefing questions. Fig.2 shows a snapshot of the software.

The Dependent Variables. While the subject is playing the games, the following dependent variable is measured: NCOOP – The number of times the subject cooperates. After game playing, a set of questions is asked to the subjects in the debriefing section of the survey. The questions refer to the players as ‘Player A’, which is the first the subject plays with and is in either the neutral or moral condition, and ‘Player B’, which is the second the subject plays with and is in the other condition. From these questions, the following dependent variables are measured: HL – Classification of how human-like was the virtual human (1 – ‘totally unlike a human’ to 6 – ‘extremely like a human’); WELF – Classification of how much was the virtual human considerate of the subject’s welfare (1 – ‘never’ to 6 – ‘always’). Subjects were also asked to choose which player they preferred to play with - the neutral, moral or no preference.



Prisoner's Dilemma with Player A: Round 4 of 25

Rules

		PARTNER	
		Remain SILENT	TESTIFY against you
YOU	Remain SILENT	Both get 3 months	You: 3 years Partner: goes free
	TESTIFY against other	You: go free Partner: 3 years	Both get 1 year

See instructions again...

Player A: 0 years 6 months

Last Action: TESTIFIED against you
Last Outcome: went free

You: 3 years 6 months

Last Action: Remained SILENT
Last Outcome: 3 year

Fig. 2. A snapshot of the survey software in the game phase.

The Hypotheses. The hypotheses we set forth for this experiment are: H1 – The subject will cooperate more with the moral virtual human than with the neutral; H2 – Subjects will perceive the moral virtual human as being more human-like than the neutral; H3 – Subjects will perceive the moral virtual human to be more considerate of the subjects’ welfare than the neutral.

Participants. Twenty-eight (28) subjects were recruited at the University of Southern California and related institutions. Subjects’ were, on average, 25.2 years of age; 42.9% were males; and, all had at least college-level education but, in diverse areas.

3 Results

The dependent *t* test was used to compare means for the dependent variables in the moral and neutral conditions. Table 1 shows descriptive statistics and the results of the *t* test. The results show that hypotheses H1 and H2 are accepted, whereas H3 is not. Subjects also self-reported that they preferred to play against the moral virtual human: fifteen subjects (53.57%) preferred the moral agent; nine subjects (32.14%) preferred the neutral agent; and four subjects (14.29%) had no preference.

Table 1. Descriptive statistics and dependent *t* test for the following dependent variables (*df* = 27): NCOOP, HL and WELF.

Vars	*Moral*		*Neutral*		*Diff.*	*Diff.*	*t*	*Sig.*	*r*
	Mean	*SD*	*Mean*	*SD*	*Means*	*SE*		*2-sd*	
NCOOP*	16.571	6.563	12.893	7.320	3.679	1.402	2.624	0.014	0.451
HL*	4.36	1.224	3.32	1.188	1.036	0.358	2.892	0.007	0.486
WELF	3.96	1.453	3.25	1.602	0.714	0.496	1.441	0.161	0.267

* Significant difference, $p < 0.05$

4 Discussion

The results suggest that people cooperate more with virtual humans if these express moral emotions. This is in line with Frank’s view that participants in social dilemmas look for contextual cues in their trading partners that they are likely to cooperate [1]. One such cue is the expression of moral emotions. As to why this cue works in our case, we look at Keltner & Kring social-functional characterization of emotions [13]. Accordingly, the display of emotions serves three functions: *informative*, signaling information about feelings and intentions to the interaction partner; *evocative*, eliciting complementary or similar empathic emotions in others; *incentive*, to reinforce, by reward or punishment, another’s individual social behavior within ongoing interactions. So, under this view, the expression of moral emotions in the virtual human is likely to be promoting cooperation because: it is informative of its willingness to engage in cooperative behavior; and, it is providing incentive for mutual cooperation through appropriate facial feedback. Regarding the evocative role of emotions, our results do not clarify whether subjects perceive the virtual human to actually be experiencing the moral emotions it expresses and whether any complimentary or empathic emotion is actually being experienced by the subjects.

The results also show that the moral virtual human is perceived as being more human-like than the neutral virtual human. This could have led to a sense of closer psychological distance between subject and moral virtual human. Psychological distance, in turn, is argued to influence the establishment of bonds of sympathy between interacting partners which, in turn, positively influences the potential for cooperation [14]. The argument here is that because subjects perceive the virtual human to be more human-like they are more likely to be sympathetic towards it and, thus, more likely to attempt cooperation. This might have also been the reason why subjects tended to prefer playing the game with the moral virtual human.

The results do not show a statistically significant difference in the perception of consideration for the subject’s welfare between the neutral and moral virtual humans. Nevertheless, subjects did cooperate more with the moral virtual human. This suggests that the expression of moral emotions could be having an unconscious influence on subjects’ decision-making and so, even though subjects were cooperating more with the moral virtual human, they were not conscious of it. This would be in line with Damasio's account of the influence of emotions in human decision-making at an unconscious level [15] and with Reeves & Nass [8] perspective that humans unconsciously treat interactions with the media (in our case, virtual humans) in the same way as with humans. The result, however, did not generalize to all subjects, as many referred explicitly to emotions (or some aspect of it) as the reason they preferred the moral virtual human to the neutral one.

In general, the results emphasize the importance of embodiment in virtual agents to the emergence of cooperation with humans. Effectively, in our study, even though the action policies were the same in both conditions, subjects cooperated more with the moral virtual human. Furthermore, this work has only begun to explore the many ways in which embodiment can play a role in the emergence of cooperation. Two promising lines of future work are the building of sympathetic bonds and rapport [16] which rely heavily on embodiment and that, in human-human interactions, contribute to the emergence of cooperation.

References

1. Frank, R.: Introducing Moral Emotions into Models of Rational Choice. In: Manstead, A. S., Frijda, N., Fischer, A. (eds) Feelings and Emotions: The Amsterdam Symposium, pp. 422--440. Cambridge University Press (2004)
2. Haidt, J.: The Moral Emotions. In: Davidson, R. J., Scherer K. R., Goldsmith, H. H. (eds.) Handbook of Affective Sciences, pp. 852--870. Oxford University Press (2003)
3. Lowenstein, G., Lerner, J.: The Role of Affect in Decision Making. In: Davidson, R. J., Scherer, K. R., Goldsmith, H. H. (eds.) Handbook of Affective Sciences, pp. 619--642. Oxford University Press (2003)
4. Frank, R.: Cooperation through emotional commitment. In: Hesse, R. (ed.), Evolution and the capacity for commitment, pp. 57--76, New York, NY: Russell Sage (2001)
5. Gratch J., Rickel, J., Andre, E., Cassell, J., Petajan, E., Badler, N.: Creating Interactive Virtual Humans: Some Assembly Required. IEEE Intellig. Systems, 17(4):54--63 (2002)
6. Sabater, J., Sierra, C.: Review on computational trust and reputation models. In: Artificial Intelligence Review, 24:33--60 (2005)
7. Morris, M., Keltner, D.: How Emotions Work: The Social Functions of Emotional Expression in Negotiations. In: Research in Organizational Behaviour, 22:1--50 (2000)
8. Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. University of Chicago Press (1996)
9. Kramer N.: Social Effects of Virtual Assistants. A Review of Empirical Results with Regard to Communication. *Intelligent Virtual Agents 2008*, pp.507-508 (2008)
10. Poundstone W.: Prisoner's Dilemma. Anchor Books (1992)
11. Axelrod R.: The Evolution of Cooperation. Basic Books (1984)
12. de Melo, C., Paiva, A.: Multimodal Expression in Virtual Humans. Computer Animation and Virtual Worlds, 17(3-4):239--248 (2006)
13. Keltner, D., Kring, A. Emotion, Social Function, and Psychopathology. Review of General Psychology, 2(3):320--342 (1998)
14. Sally, D.: A general theory of sympathy, mind-reading, and social interaction, with an application to the prisoner's dilemma. Social Science Information, 39(4):567--623 (2000)
15. Damasio A.: Descarte's Error: Emotion, Reason, and the Human Brain. G.P. Putnan's Sons (1994)
16. Capella, J.: On defining conversational coordination and rapport. In: Psychological Inquiry, 1(4):303—305 (1990)