The Impact of Emotion Displays in Embodied Agents on Emergence of Cooperation with People

Authors: Celso M. de Melo (demelo@ict.usc.edu, corresponding author), Peter Carnevale (peter.carnevale@marshall.usc.edu) and Jonathan Gratch (gratch@ict.usc.edu)

Affiliation of 1st and 3rd Authors: Institute for Creative Technologies, University of Southern California, 12015 Waterfront Drive, Building #4 Playa Vista, CA 90094-2536, USA

Affiliation of 2nd Author: USC Marshall School of Business, Los Angeles, CA 90089-0808, USA

Abstract

Acknowledging the social functions of emotion in people, there has been growing interest in the interpersonal effect of emotion on cooperation in social dilemmas. This article explores whether and how facial displays of emotion in embodied agents impact cooperation with human users. The article describes an experiment where participants play the iterated prisoner's dilemma against two different agents that play the same strategy (tit-for-tat), but communicate different goal orientations (cooperative vs. individualistic) through their patterns of facial displays. The results show that participants are sensitive to differences in the emotion displays and cooperate significantly more with the cooperative agent. The results also reveal that cooperation rates are only significantly different when people play first with the individualistic agent. This is in line with the well-known black-hat/white-hat effect from the negotiation literature. However, this study emphasizes that people can discern a cooperator (white-hat) from a non-cooperator (black-hat) based only on emotion displays. We propose that people are able to identify the cooperator by inferring from the emotion displays, the agent's goals. We refer to this as reverse appraisal, as it reverses the usual process in which appraising relevant events with respect to one's goals leads to specific emotion displays. We discuss implications for designing human-computer interfaces and understanding human-human interaction.

1. Introduction

A growing number of studies have explored emotion in embodied agents to enhance interaction with computers (Beale & Creed, 2009). Acknowledging the idea that people interact with computers in a social manner just like with other people (Nass, Steuer, & Tauber, 1994; Reeves & Nass, 1996), researchers have attempted to create agents that display emotions in ways that are consistent with displays people show in daily interaction. However, the current focus of research has been on showing that emotion *can* enhance interaction (Beale & Creed, 2009; Dehn & Van Mulken, 2000), rather than on understanding the mechanisms by which emotion influences human-agent interaction. Thus, many studies focus on simple comparisons between agents that display emotions when compared to agents that do not (Lester et al., 1997; Maldonado et al., 2005; Prendinger, Mayer, Mori, & Ishizuka, 2003; Klein, Moon, & Picard, 2002; Hone, 2006; Liu & Picard, 2005; Lim & Aylett, 2007); some studies compare agents that display consistent versus inconsistent emotions (Berry, Butler, & De Rosis, 2005; Creed & Beale, 2008), and naturally conclude that people prefer agents that display consistent emotions; and, the few studies that compare agents that express different emotions compare simple aspects of emotion and do not frame the results within a broad theory of emotion: Gong (2007) shows that people prefer an agent that displays positive emotions to one that displays negative emotions, independently of context; and, Brave, Nass and Hutchinson (2005) show that people prefer agents that display other-empathetic emotions to agents that displays self-empathetic emotions. As a result, a crude view of the impact of emotion in embodied agents emerges, which we refer to as the *affective persona effect*, which argues that the mere presence of consistent emotions in embodied agents is sufficient to improve human-machine interaction. This view can be seen as a straightforward extension of the *persona* *effect* (Lester et al., 1997; Van Mulken, André, & Muller, 1998) – which argues that the mere presence of embodied agents is sufficient to enhance human-machine interaction – to the case of agents that display emotions.

In this article we are interested on the impact emotion in embodied agents can have on people's decision-making and, in particular, on emergence of cooperation in human-agent interaction. In line with the affective persona effect, we have shown in the past that people cooperate more with an embodied agent that displays emotions than with an agent that does not (de Melo, Zheng, & Gratch, 2009). In this study, participants played a social dilemma game with agents that followed the same strategy to choose their actions – tit-for-tat – but, one displayed emotions consistent with a goal of mutual cooperation. We referred to this agent as the *cooperative* agent and the emotion displays were as follows: when both players cooperated, it expressed gratitude (as the outcome was appraised to be positive for the self and, the participant was appraised to have contributed to it); when the agent cooperated but the participant defected, it expressed anger (as the outcome was negative for the self and the participant was blamed for it); when the agent defected and the participant cooperated, it expressed shame (as the outcome was negative for the participant and the agent blamed itself for it); when both defected, it expressed sadness (as this outcome was negative for both). The results showed that participants cooperate significantly more with the cooperative agent than the control agent (that showed no emotion). Moreover, participants reported preferring to play with the cooperative agent and perceiving it to be more human-like than the control agent. However, we were not satisfied with the view that the mere presence of emotion in the cooperative agent was sufficient to explain the results. We believe that context is crucial for interpreting emotions and the fact that the facial displays in the cooperative agent are compatible with a goal of mutual cooperation is critical for the effect to occur. Therefore, in this article we present a new experiment that shows the insufficiency of the affective persona effect to explain the impact of emotion on cooperation and, we look instead to theories of the social functions of emotion to understand the role of emotion on emergence of cooperation in human-agent interaction.

Theories of the social functions of emotion (Frijda & Mesquita, 1994; Keltner & Haidt, 1999; Keltner & Kring, 1998; Morris & Keltner, 2000) argue that emotions convey information about one's goals, desires and beliefs to others and, thus, help regulate social interaction. In particular, this view has led to the idea that displays of emotion can be used to identify cooperators in social dilemmas. Frequently people are faced with situations where they must choose between pursuing their own self-interest and collect a short-term reward or rely on another person for mutual cooperation and maximize joint long-term reward (Frank, 2004; De Cremer, 1999; Kollock, 1998). In these cases, it is valuable, from an adaptive point of view, to be able to detect how likely the other is to cooperate (Dawkins, 1976; Frank, 1988; Hamilton, 1964; Trivers, 1971). Nonverbal displays have been argued to be an important cue in this detection process (Boone & Buck, 2003). In particular, there has been a lot of empirical research on the impact of facial displays of emotion on emergence of cooperation: many studies show that cooperative individuals display higher levels of positive emotion than non-cooperators (Scharlemann, Eckel, Kacelnik, & Wilson, 2001; Mehu, Grammer & Dunbar, 2001); Krumhuber, Manstead and Kappas (2007) show that the dynamics of facial displays are also relevant for the perception of trustworthiness; Chapman, Kim, Susskind and Anderson (2009) show that disgust can also reveal pro-social tendencies in certain situations; and, Schug, Matsumoto, Horita, Yamagishi, & Bonnet (2010) argue that, aside from positive displays of emotion, cooperators can also be identified from negative displays of emotion.

One theory inspired on the social-functions view of emotion is based on appraisal theories of emotion. In appraisal theories (Ellsworth & Scherer, 2003), emotion displays arise from cognitive appraisal of events with respect to an agent’s goals, desires and beliefs (e.g., is this event congruent with my goals? Who is responsible for this event?). According to the pattern of appraisals that occurs, different emotions are experienced and displayed. Now, since displays reflect the agent’s intentions through the appraisal process, it is also plausible to ask whether people can infer from emotion displays the agent’s goals by reversing the appraisal mechanism. We refer to this theory as the *reverse appraisal* theory. Empirical evidence is still scarce but in a recent study Hareli and Hess (2009) show that people can, from expressed emotion, make inferences about the character of the person displaying emotion. So, for instance, a person who reacted with anger to blame was perceived as being more aggressive, self-confident but also as less warm and gentle than a person who reacted with sadness. In this article we explore reverse appraisal as a possible mechanism by which facial displays in embodied agents impact cooperation with people.

We describe a new experiment where participants play the iterated prisoner’s dilemma with two agents that follow the same strategy to choose their actions but have different emotion display policies. The *cooperative* agent remains unchanged from our aforementioned previous experiment. However, instead of comparing it to a control agent that has no emotions, we compare it to an *individualistic* agent which emotions reflect pure self-interest, i.e., the goal of maximizing its own points independently of the value of the outcome to the participant. Thus, when this agent defects and the participant cooperates, it expresses joy (as this event is appraised to be very positive); when the participant defects and the agent cooperates, it expresses sadness (as this is the worst event for the self); and so on. According to the affective persona effect, we should expect no difference in terms of participant cooperation with the agents, as both agents express consistent emotions (even though consistent with different goals). However, according to the reverse appraisal theory, we should expect participants to infer the agent’s goals from the emotion displays, decide based on those inferences and, finally, cooperate more with the cooperative agent. Thus, our hypothesis (H1) is that people will cooperate significantly more with the cooperative agent than the individualistic agent.

Finally, ordering effects have been reported in the decision-making literature when people play in sequence with a cooperator and a non-cooperator (Hilty & Carnevale, 1993; Hartford & Solomon, 1967; Bixenstine & Wilson, 1963). In particular, a well-studied contrast effect in the negotiation literature is known as the *black-hat/white-hat* (or *bad-cop/good-cop*) effect (Hilty & Carnevale, 1993). In bilateral negotiation, Hilty and Carnevale (1993) showed that playing a first game with an opponent with a competitive stance (black-hat) followed by a second game with an opponent with a cooperative stance (white-hat) is more effective in reducing distance to agreement than any other pairing of the black-hat and white-hat opponents (white-hat/white-hat, white-hat/black-hat and black-hat/black-hat). In the prisoner's dilemma, Harford and Solomon (1967) found that a "reformed sinner" strategy (a change in behavior from less cooperation to more cooperation, which is analogous to the black-hat/white-hat strategy) elicited higher levels of cooperation than other strategies, such as the "pacifist" strategy (analogous to the white-hat/white-hat strategy). Also, Bixenstine and Wilson (1963) found that initial noncooperation followed by cooperative behavior elicited higher levels of cooperation. One explanation for the effectiveness of the black-hat/white-hat strategy relies on the dynamics of reciprocity. Reciprocity in negotiation is manifest in "matching" or strategy imitation, in which a bargainer concedes when the other concedes, or is firm when the other is perceived as firm (Pruitt & Carnevale, 1993). Whether people will match concessions, is dependent on context: if concessions are attributed to weakness, this will encourage exploitation (Deutsch, Epstein, Canavan, & Gumpert, 1967). This suggests that initial firmness may lessen the temptation to exploit and that cooperative initiatives that are extended in the context of firmness may be more likely to evoke reciprocity. Another explanation of the black-hat/white-hat effect is based on the concepts of adaptation and comparison level (Helson, 1964). Theories of adaptation propose that people become accustomed to a neutral reference point as a result of prior experience; this point then serves as a comparison for judgment of subsequent experiences. Thus, a cooperative second bargainer should be judged as more cooperative if the first bargainer was competitive rather than cooperative. This positive shift in perception of cooperativeness should, in turn, foster mutual cooperation. In this article we also explore whether contrast effects occur when people play the cooperative (white-hat) and individualistic agent (black-hat) in different orders. Our hypothesis (H2) is that participants will cooperate more with the cooperative agent, predominantly when playing with the individualistic agent first.

2. Experiment

The experiment follows a repeated-measures design where participants play 25 rounds of the iterated prisoner's dilemma with two different computational agents for a chance to win real money: the cooperative agent; and the individualistic agent. The agents differ in the way their facial displays reflect the outcome of each round. The action policy, i.e., the strategy for choosing which action to take in each round, is the same for both agents.

2.1 Game

Following the approach by Kiesler, Waters and Sproull (1996), the prisoner's dilemma game was recast as an investment game and described as follows to the participants: "You are going to play a two-player investment game. You can invest in one of two projects: Project Green and Project Blue. However, how many points you get is contingent on which project the other player invests in. So, if you both invest in Project Green, then each gets 5 points. If you choose Project Green but the other player chooses Project Blue, then you get

3 and the other player gets 7 points. If, on the other hand, you choose Project Blue and the other player chooses Project Green, then you get 7 and the other player gets 3 points. A fourth possibility is that you both choose Project Blue, in which case both get 4 points". There are, therefore, two possible actions in each round: *Project Green* (or cooperation); and *Project Blue* (or defection). Table 1 summarizes the payoff matrix. The participant is told that there is no communication between the players before choosing an action. Moreover, the participant is told that the agent makes its decision without knowledge of what the participant's choice in that round is. *After* the round is over, the action each chose is made available to both players and the outcome of the round, i.e., the number of points each player got, is also shown. The experiment is fully implemented in software and a snapshot is shown in Figure 1.

Table 1. Payoff matrix for the investment game.

		Agent	
		Project Green	Project Blue
Participant	Project Green	Agent: 5 pts Participant: 5 pts	Agent: 7 pts Participant: 3 pts
	Project Blue	Agent: 3 pts Participant: 7 pts	Agent: 4 pts Participant: 4 pts

"Figure 1 here"

2.2 Action Policy

Agents in both conditions play the same action policy, i.e., they follow the same strategy to choose their actions. The policy is 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 has been argued to strike the right balance of punishment and reward with respect to the opponent's previous actions (Axelrod, 1984). So, the action policy used in our experiment is as follows: (a) in rounds 1 to 5, the agent plays the following fixed sequence: cooperation, cooperation, defection, defection, cooperation; (b) in rounds 6 to 25, the agent plays pure tit-for-tat. The rationale for the sequence in the first five rounds is to make it harder for participants to learn the agents' strategy and to allow participants to experience a variety of facial displays from the start.

2.3 Conditions

There are two conditions in this experiment: the *cooperative* agent; and the *individualistic* agent. Both agents follow the same action policy but differ in their facial display policies. The facial display policy defines the emotion and intensity which is conveyed for each possible outcome of a round. Table 2 shows the facial displays for the cooperative agent and Table 3 for the individualistic agent. The facial displays are chosen to reflect the agents' goals in a way that is consistent with appraisal models of emotion (Ellsworth & Scherer, 2003). The cooperative agent has the goal of reaching mutual cooperation. Thus, when both players cooperate, it will express gratitude (with a facial display of joy), as the outcome is appraised to be positive for the self and the participant is appraised to have contributed for it; when the agent defects and the participant cooperates, it expresses shame, as the outcome is negative for the participant and the agent is responsible; when the agent cooperates and the participant defects, it expresses anger, as the outcome is negative and the participant is responsible for it; and, when both defect, it expresses sadness, as the event is negative. The individualistic agent, on the other hand, has the goal of maximizing its own points (independently of the value of the outcome for the other player). Therefore, when the agent defects and the participant cooperates, it expresses joy, as this event is appraised to be very positive; when both cooperate, it expresses nothing, as this event could be more positive; when both defect, it expresses sadness at 50%[1], as this is a negative event; when the participant defects and the agent cooperates, it expresses sadness at 100%, as this is the

[1] Expression of sadness at 50% corresponds to 50% interpolation between the neutral and sadness facial displays (shown in Figure 2).

worst event for the self. Facial displays are animated using a real-time pseudo-muscular model for the face which also simulates wrinkles and blushing (de Melo & Gratch, 2009). The facial display is shown at the end of the round, after both players have chosen their actions and the outcome is shown. Moreover, there is a 4.5 seconds waiting period before the participant is allowed to choose the action for the next round. This period allows the participant to appreciate the outcome of a round before moving to the next round. Finally, to enhance naturalness, blinking is simulated in both agents as well as subtle random motion of the neck and back.

Table 2. Facial displays (emotion and intensities) for the cooperative agent.

Cooperative Agent		*Agent*	
		Project Green	Project Blue
Participant	Project Green	Joy (100%)	Shame (100%)
	Project Blue	Anger (100%)	Sadness (100%)

Table 3. Facial displays (emotion and intensities) for the individualistic agent.

Individualistic Agent		*Agent*	
		Project Green	Project Blue
Participant	Project Green	Neutral	Joy (100%)
	Project Blue	Sadness (100%)	Sadness (50%)

The condition order is randomized while making sure that 50% of the participants experience one order and the remaining 50% the other. Two different bodies are used: *Michael* and *Daniel*. These bodies are shown in Figure 2 as well as their respective facial displays. Bodies are assigned to each condition in random order and agents are referred to by the names of their bodies throughout the experiment.

“Figure 2 here”

To validate the facial displays, a pre-study was conducted where participants were asked to classify, from 1 (meaning 'not at all') to 5 (meaning 'very much'), how much each of the displays conveys joy, sadness, shame and anger. Images of the displays and questions were presented in random order. Twenty-two participants were recruited just for this study from the same participant pool as the main experiment (described below). The results are shown in Table 4. A *repeated-measures ANOVA* was used to compare the means for perceived emotion in each display. Significant differences were found for all displays except, as expected, for the neutral case. Moreover, pairwise comparisons of the perception of the real emotion with respect to perception of the other emotions were all significant in favor of the real emotion, with one exception: displays of shame were also significantly perceived as displays of sadness. This is not a problem since it is usually agreed that shame occurs upon the occurrence of a negative event, thus causing sadness, plus the attribution of blame to the self (Ortony, Clore & Collins, 1988).

Table 4. Classification of the facial displays with respect to perception of joy, sadness, shame and anger. Scale goes from 1 (meaning 'not at all') to 5 (meaning 'very much').

	Perceived Emotion			
	Joy	*Sadness*	*Shame*	*Anger*
Real Emotion	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
		Michael		
Neutral	1.86 (.941)	1.86 (1.037)	1.91 (1.065)	1.68 (.945)
Joy*	4.05 (.899)	1.18 (.501)	1.23 (.528)	1.41 (1.098)
Sadness*	1.27 (.703)	4.09 (1.019)	2.77 (1.478)	1.50 (.859)
Shame*	1.32 (.716)	3.59 (1.182)	3.55 (1.371)	1.45 (.858)
Anger*	1.36 (.727)	1.95 (1.046)	1.32 (.646)	4.32 (1.211)
		Daniel		
Neutral	1.55 (1.057)	1.73 (.935)	1.68 (.894)	2.18 (1.259)
Joy*	3.77 (1.020)	1.18 (.501)	1.23 (.528)	1.14 (.468)
Sadness*	1.41 (.854)	3.68 (1.492)	2.73 (1.386)	1.50 (.740)
Shame*	1.32 (.780)	3.77 (1.412)	3.86 (1.356)	1.41 (.734)
Anger*	1.27 (.703)	1.82 (1.332)	1.55 (1.011)	4.27 (1.420)

* Significant difference between means based on *repeated-measures ANOVA*, *p* < .05

2.4 Measures

During game-play, we save information regarding whether the participant cooperated in each round. This is our main behavioral measure. After playing with each agent, we ask how human-like was the agent (scale goes from 1 - 'not at all' to 6 - 'very much'). After the game is over, to try to understand how the agents are being interpreted, the participant is asked to classify each agent according to the Person-Perception scale (Bente, Feist, & Elder, 1996) which consists of 33 bipolar pairs of adjectives: dislikable-likable; cruel-kind; unfriendly-friendly; cold-warm; unreliable-reliable; relaxed-tense; detached-involved; rude-polite; dishonest-honest; unpleasant-pleasant; naïve-sophisticated; unapproachable-inviting; passive-active; aloof-compassionate; non-threatening-threatening; not cool-cool; unintelligent-intelligent; cold-sensitive; sleepy-alert; proud-humble; unsympathetic-sympathetic; shy-self-confident; callous-tender; permissive-stern; cheerful-sad; modest-arrogant; not conceited-conceited; weak-strong; mature-immature; noisy-quiet; nervous-calm; soft-tough; acquiescent-emancipated. In this scale items are rated on a 7-point scale (e.g., 1-'dislikable' to 7-'likable'). Finally, participants are asked 'Which agent did you prefer to play with?' as well as two exploratory classification questions (scale goes from 1-'never' to 6-'always'), where agents are actually referred to by the names of their bodies:

- How considerate of your welfare was the <cooperative/individualistic agent>?
- How much would you trust the <cooperative/individualistic agent>?

2.5 Participants

Fifty-one participants were recruited at the University of Southern California Marshall School of Business. Average age was 21.0 years. Gender distribution was as follows: *males*, 45.1%; *females*, 54.9%. Most participants were undergraduate students (96.9%) majoring in business (86.3%). Most were also originally from the United States (84.3%).

The incentive to participate follows standard practice in experimental economics (Hertwig & Ortmann, 2001): first, participants were given credit for their participation in this experiment; second, with respect to their goal in the game, participants were instructed to earn as many points as possible, as the total amount of points would increase their chance of winning a lottery for $100.

3. Results

3.1 Cooperation

To understand how people cooperate with the agents in each condition, the following variables were defined:

- Coop.All – cooperation rate over all rounds;
- Coop.AgC – cooperation rate when the agent cooperated in the previous round;
- Coop.AgD – cooperation rate when the agent defected in the previous round.

The *Kolmogorov-Smirnov* test was applied to all these variables to test for their normality and all were found to be significantly non-normal. Therefore, the *Wilcoxon signed ranks* test is used to compare means between conditions. The results, shown in Table 5, indicate that people cooperate significantly more with the cooperative agent (*M*=.37, *SD*=.28) than the individualistic agent (*M*=.27, *SD*=.23; *p*<.05, *r*=.320). Thus, our hypothesis H1 is confirmed. The results also suggest that this difference in cooperation is particularly salient following a defection by the agent.

Table 5. Descriptive statistics and significance levels for percentage of cooperation.

Variables	*Cooperative*		*Individualistic*		*Sig.*	*|r|*
	Mean	*SD*	*Mean*	*SD*	*2-sd*	
Coop.All*	.366	.279	.272	.231	.022	.320
Coop.AgC	.397	.319	.339	.288	.262	ns
Coop.AgD*	.297	.256	.203	.197	.022	.320

* Significant difference, p < 0.05

Figure 3 shows how cooperation rate (Coop.All) evolves with each round. The graph shows that people start cooperating less with the individualistic agent as early as the 3rd round. Even though both agents defect in rounds 3 and 4 (see the 'Experiment' section), participants cooperate much less with the individualistic agent in round 5. After the agents cooperate in rounds 5 and 6, people seem to attempt cooperation again in round 7 with the individualistic agent but, from then on, again consistently cooperate less with the individualistic agent.

"Figure 3 here"

As discussed in the Introduction, studies have shown that when people engage in sequence with a cooperator and a non-cooperator in a social dilemma, the order of interaction can have an impact on level of cooperation (Hilty & Carnevale, 1993; Harford & Solomon, 1967; Bixenstine & Wilson, 1963). To explore whether order is having an effect in cooperation, Table 6 shows cooperation rates for each condition order. The results are clear and reveal that the effect described above (Table 5) is being driven by the order individualistic agent first, cooperative agent second. Effectively, cooperation does not differ significantly between conditions when participants play with the cooperative agent first. Therefore, our hypothesis H2 is also confirmed.

Table 6. Descriptive statistics and significance levels for percentage of cooperation by condition order.

Variables	*Cooperative*		*Individualistic*		*Sig.*	|*r*|
	Mean	*SD*	*Mean*	*SD*	*2-sd*	
	Cooperative → Individualistic (N=26)					
Coop.All	.345	.260	.309	.261	.572	ns
Coop.AgC	.380	.301	.367	.314	.897	ns

Coop.AgD	.267	.203	.207	.203	.232	ns
Individualistic → Cooperative (N=25)						
Coop.All*	.389	.302	.234	.192	.016	.484
Coop.AgC	.414	.342	.310	.260	.159	ns
Coop.AgD*	.329	.303	.199	.195	.064	.370

* Significant difference, p < 0.05

To understand whether people's decision-making was reflecting only facial displays, as opposed to facial displays *and* round outcome, we compared percentage of cooperation for the same display between conditions. Table 7 shows these results. Significance values are calculated using the *Wilcoxon signed ranks* test. The results show that for the same display of joy, participants cooperate significantly more with the cooperative agent than the individualistic agent. The results also show that, once again, this effect is driven by the order individualistic agent first, cooperative agent second.

Table 7. Comparison of cooperation rates for the same facial display between conditions. Joy occurs when there is mutual cooperation in the cooperative condition and when the agent defects and the participant cooperates in the individualistic condition. Sadness occurs when there is mutual defection in the cooperative condition and when the participant defects and the agent cooperates in the individualistic condition.

Variables	*Cooperative*		*Individualistic*		*Sig.*	$\lvert r \rvert$
	Mean	*SD*	*Mean*	*SD*	*2-sd*	
All Orders (N=51)						
Joy*	.417	.417	.223	.282	.008	.371
Sadness	.290	.310	.245	.218	.484	ns
Cooperative → Individualistic (N=26)						
Joy	.392	.387	.245	.317	.150	ns
Sadness	.300	.321	.259	.246	.487	ns
Individualistic → Cooperative (N=25)						
Joy*	.442	.452	.200	.246	.016	.481
Sadness	.278	.305	.232	.190	.414	ns

* Significant difference, p < 0.05

Since there is evidence that people form judgments of people based only on appearance (Willis & Todorov, 2006), we wanted to make sure that the body was not a confounding factor in our experiment. Thus, we compared percentage of cooperation between the two agent bodies used in the experiment. It was found that there was *no* significant difference in cooperation between Michael (*M*=.33, *SD*=.26) and Daniel (*M*=.31, *SD*=.26; *p*>.05). Significance level is calculated using the *Wilcoxon signed ranks* test.

3.2 Agent Characterization

Principal component analysis (varimax rotation, scree-test) on the Person-Perception scale revealed three factors consistent with the literature (Bente et al., 1996): *evaluation*, explains 33.1% of the variance (Cronbach's Alpha = .962) with main loading factors of friendly-unfriendly, kind-cruel and sympathetic-unsympathetic; *potency* (or *power*), explains 17.5% of the variance (Cronbach's Alpha = .902) with main loading factors of emancipated-acquiescent, tough-soft and arrogant-modest; *activity,* explains 8.0% of the variance (Cronbach's Alpha = .762) with main loading factors of active-passive, involved-detached and alert-sleepy. These three factors were calculated for both conditions and the means compared using the *dependent t test* (since the *Kolmogorov-Smirnov* test was not significant). The results are shown in Table 8. When collapsing across condition order, the results reveal two non-significant trends: (a) participants perceive the individualistic agent to be more powerful than the cooperative agent; (b) people perceive the cooperative agent to be more active than the individualistic agent. However, when we consider only the order where participants play the cooperative agent first, then these trends become significant.

Table 8. Descriptive statistics and significance levels for Person-Perception scale.

Variables	*Cooperative*		*Individualistic*		*Sig.*	*r*
	Mean	*SD*	*Mean*	*SD*	*2-sd*	
			All Orders (N=51)			
Evaluation	5.01	1.749	4.74	1.605	.461	ns

Potency/Power	5.26	1.539	5.81	1.263	.086	ns
Activity	3.88	1.485	3.50	.937	.163	ns
Cooperative → Individualistic (N=26)						
Evaluation	5.110	1.673	4.64	1.425	.293	ns
Potency/Power*	4.90	1.266	5.71	1.377	.048	.385
Activity*	4.04	1.443	3.16	.728	.005	.521
Individualistic → Cooperative (N=25)						
Evaluation	4.90	1.853	4.84	1.797	.913	ns
Potency/Power	5.64	1.724	5.91	1.538	.588	ns
Activity	3.71	1.538	3.84	1.016	.228	ns

* Significant difference, $p < 0.05$

The results for the human-likeness and post-game questions are shown in Table 9. Significance values are calculated using the *Wilcoxon signed ranks* test. The results show that there are no differences in terms of perception of human-likeness between the agents. The results also show that people perceive the cooperative agent to be marginally significantly more trustworthy than the individualistic agent but, there was no significant difference in perception of consideration of the participant’s welfare.

Table 9. Descriptive statistics and significance levels for post-game and human-likeness questions.

Variables	*Cooperative*		*Individualistic*		*Sig. 2-sd*	|*r*|
	Mean	*SD*	*Mean*	*SD*		
All Orders (N=51)						
Human-like?	2.90	1.233	2.98	1.204	.943	ns
Considers your Welfare?	3.06	1.683	2.82	1.466	.327	ns
Trustworthy?	3.04	1.665	2.56	1.280	.065	.260
Cooperative → Individualistic (N=26)						
Human-like?	2.85	1.287	3.04	1.113	.531	ns
Considers your Welfare?	3.23	1.608	3.23	1.478	.972	ns
Trustworthy?	3.15	1.592	2.65	1.231	.102	ns
Individualistic → Cooperative (N=25)						
Human-like?	2.96	1.197	2.92	1.316	.623	ns
Considers your Welfare?	2.88	1.345	2.38	1.345	.187	ns
Trustworthy?	2.92	1.767	2.46	1.351	.262	ns

Finally, results for agent preference are shown in Table 10. Significance levels are calculated using the *Chi-square* test. When collapsing across orders or playing with the individualistic agent first, people do not show a significant agent preference. However, when playing with the cooperative agent first, there is a trend for preferring to play with the cooperative agent.

Table 10. Results for agent preference.

Order	*Cooperative*	*Individualistic*	*All Same*	*Sig.*
All Orders	45.1%	29.4%	25.5%	.193
Cooperative → Individualistic	53.9%	26.9%	19.2%	.076
Individualistic → Cooperative	36.0%	32.0%	32.0%	.961

4. Discussion

The results show that people cooperate more with the cooperative agent than with the individualistic agent. This finding is in line with the view that, in social dilemmas, people look for cues in their trading partners that they might be willing to cooperate before engaging in cooperation themselves (Frank, 2004; Boone & Buck, 2003). What our results suggest is that people also care and look for these cues when engaged in a social dilemma with embodied agents. But, why are people cooperating more with the cooperative agent than the individualistic agent? Let’s start by excluding the affective persona effect as a possible explanation for the results. Recall this theory argues that embodied agents that express consistent emotions enhance human-machine interaction. In our previous study (de Melo et al., 2009), and in line with this theory, we showed that people cooperated significantly more with the cooperative agent than with a control agent that expressed no emotion. However, though promising, the results did not prove that the agent needed to have “cooperative” emotions in order to promote cooperation. The argument is that the mere fact the agent had emotions, “cooperative” or not, led to increased engagement and this alone was sufficient to explain the increase in cooperation. However, this explanation cannot apply to the current experiment as both the cooperative and individualistic agents display emotions. Moreover, in the previous study people perceived the cooperative agent to be more human-like than the control agent. However, in this study both agents were perceived to be equally human-like (Table 9) and, thus, we can also exclude the explanation that people cooperate with the most human-like of the agents.

We argue people are using the facial displays conveyed by the agents to learn about the agents' goals and, then, act accordingly. The social-functions view of emotion (Frijda & Mesquita, 1994; Keltner & Haidt, 1999; Keltner & Kring, 1998; Morris & Keltner, 2000) argues that the display of emotions can serve an informative function, signaling information about feelings and intentions to the interaction partner. Our argument, then, is that the agent's emotion displays convey information people use to infer about the agent's propensity to cooperate. As to the mechanism by which people make these inferences, we argue people reverse the usual emotion appraisal mechanism and infer, from emotion displays, how is the agent appraising the outcomes (e.g., is the outcome positive? Does it blame me for the outcome?) and, thus, what its goals are. For instance, if after the participant cooperates and the agent defects, the agent displays shame (as in the case of the cooperative agent), then the participant can infer that this outcome is appraised as negative by the agent and, moreover, that the agent believes itself to be at blame. However, if for the same actions, the agent displays joy (as in the case of the individualistic agent), then the participant can infer that the agent finds the outcome positive and, thus, is likely to keep defecting. The proposal, thus, explains how people infer that the cooperative agent is interested in reaching mutual cooperation and that the individualistic agent is not. To understand how these inferences materialize into different cooperation rates, we need now to consider the order participants play the agents.

The results show that people always tend to cooperate more with the cooperative agent than the individualistic agent but, this trend only becomes significant when they play the individualistic agent first. This finding is in line with the well-studied contrast effect known as the black-hat/white-hat (or bad-cop/good-cop) effect (Hilty & Carnevale, 1993; Harford & Solomon, 1967; Bixenstine & Wilson, 1963). When applied to our study, this means the cooperative agent is interpreted as the white-hat and the individualistic agent as the black-hat. However, whereas in the classical studies a cooperative or competitive stance is signaled through different levels of concession in the offers, in the current study a cooperative or competitive stance is signaled only through emotion displays. The argument then is that when participants face a tough individualistic agent in the first game, they'll be less likely to attempt exploitation in the second game and reciprocate to a more (expressively) cooperative agent. Effectively, the results on cooperation rate suggest that participants are likely exploiting the cooperative agent when they play with it first but, on the other hand, when they play the individualistic agent first, they reciprocate to the cooperative agent. In summary, we argue people use reverse appraisal to identify, from emotion displays, the cooperator (white-hat) and the non-cooperator (black-hat); then, the black-hat/white-hat contrast effect explains why participants cooperate significantly more with the cooperative agent only after playing first with the individualistic agent.

The results on the agent characterization measures provide further support for our proposal. In line with the idea that one agent is perceived as the cooperator and the other is not, the post-game classification questions (Table 9) reveal a tendency for the cooperative agent to be perceived as more trustworthy than the individualistic agent. The results on the Person-Perception scale (Table 8) also suggest people perceive the individualistic agent to be more powerful and less active/responsive than the cooperative agent. The individualistic agent is likely perceived as more powerful because its facial displays reflect only its own utility and not that of the participant and, thus, the agent is perceived as not caring about mutual cooperation, which can be viewed as an expression of power. The result on activity, in turn, likely reflects two things: that the individualistic agent is perceived to be less responsive to the participants' attempts of cooperation; and, the simplicity of the self-centered emotions of the individualistic agent when compared to the more complex other-oriented emotions of the cooperative agent. Effectively, it has been argued that social-oriented emotions such as shame (displayed by the cooperative agent) are more complex than joy and sadness (displayed by the individualistic agent) and, accordingly, tend to evolve later in life (Lewis, 2008). Overall the characterization of the agents according to power and activity is compatible with the perception that the cooperative agent cares more about the participant's interests than the individualistic agent.

However, despite the fact that the trends are in the right directions, the results for the subjective measures tend not to be significant (Tables 8, 9 and 10). We believe this reflects that the agents' emotion displays are impacting people's decision-making at an *unconscious* level. Anecdotally, in our debriefing sessions, it was not uncommon for participants, even though confirming they noticed the emotions in the agents, to state that they were not being influenced by them when deciding what to do. (This is, of course, in contrast with what the results on cooperation rate actually show.) Effectively, emotions had already been argued to influence decision-making at an unconscious level by Damasio (1994); Reeves and Nass (1996) also suggest that people treat, unconsciously, interactions with the media (in our case, embodied agents) in the same way as with real humans. Notice this would not invalidate our explanation based on reverse appraisal, as appraisals can be more cognitive or occur subcortically and automatically (Ellsworth & Scherer, 2003). Two interesting exceptions, nevertheless, do occur, but only for the white-hat/black-hat order: (1) participants perceive the individualistic agent to be significantly less active and more powerful than the cooperative agent; (2) participants (marginally) significantly prefer to play with the cooperative agent. These exceptions might occur because the white-hat/black-hat order emphasizes the selfishness of the individualistic agent’s displays (in contrast to the selflessness of the cooperative agent’s displays). This focus of attention on the contrast might, then, lead participants to become more conscious of the differences between the agents and, in turn, this is reflected in the self-reported measures. Overall, further research will be necessary to clarify which aspects of the participants’ decision-making is occurring at an unconscious level in this experimental paradigm.

The results in this paper suggest important consequences for the design of embodied agents. First, despite the large amount of empirical studies, it is still not clear whether embodied agents that display emotions can enhance human-computer interaction (Beale & Creed, 2009). This article adds evidence that embodied agents that express emotions can influence the emergence of cooperation with people. Second, our results emphasize that, contrary to the predictions of the affective persona effect theory, this effect depends on the nature of the emotions being expressed. People can differentiate between two agents that display consistent emotions, if these displays are consistent with different goals. Effectively, in our study both agents are expressing consistent emotions, but, participants selectively cooperate more with the cooperative agent. Third, we propose that participants are interpreting the agents’ emotions through a process of reverse appraisal where they infer from perceived emotion the agents’ goals, desires and beliefs. This proposal still needs further investigation and empirical evidence. Nevertheless, if this is the case, then it would mean that the mechanism defining when and which emotions the agent expresses should reflect the goals, desires and beliefs we want the user to perceive the agent to have. Consequently, computational models of appraisal theory (Marsella, Gratch, & Petta, 2010) would constitute a promising approach to synthesizing emotions in agents that people could understand.

The results are also in line with predictions from theories in the social sciences. Effectively, participants do seem to care about social cues, such as facial displays, when interacting with an agent in a social dilemma, which is in line with Frank's proposal regarding human-human interactions (Frank, 2004). Moreover, the results suggest that the social functions of emotions we see in people (Frijda & Mesquita, 1994) also carry to human-agent interactions. Altogether, the results provide further evidence that it is possible to study human-human interaction from human-agent interaction and they also reveal the potential of embodied agents as a research tool for doing basic human-human interaction research, as has already been noticed (Bente, Kramer, Petersen, & de Ruiter, 2001; Blascovich et al., 2002).

There is plenty of future work ahead. First, further alternative explanations for the effect of emotion display on cooperation rate need to be excluded: (a) the cooperative agent shows more emotions than the individualistic, as the latter expresses no emotion when both players cooperate; (b) the cooperative agent shows more distinct emotions (joy, sadness, shame and anger) than the individualistic (joy and sadness). We have begun addressing these issues in a variant of the current experiment where we compare two new versions of the cooperative and individualistic agents that express the same number and type emotions but, of course, the emotion displays are mapped differently to the dilemma's outcomes. Preliminary results show that, as expected, participants are still cooperating more with the cooperative agent and, thus, the aforementioned alternative explanations can be excluded. Second, we have already compared previously the cooperative agent with a control agent (de Melo et al., 2009), but this should also be done for the individualistic agent. Third, in the current experimental design, agents follow a variant of the widely used tit-for-tat strategy to choose their actions (Axelrod, 1984). However, this might raise the concern that, at least in the portion of the game where it is in use (rounds 6 to 25), it's not just facial displays but also the reciprocity inherent to this strategy that leads to a strong effect on cooperation rate. This issue can be addressed with a new design where tit-for-tat is replaced with a fixed strategy. Finally, we propose that participants are interpreting the agents' emotions through a process of reverse appraisal where they infer from perceived emotion the agents' goals, desires and beliefs. We have already begun collecting further evidence for this proposal in a design that attempts to show that appraisal variables (e.g., how desirable is a certain outcome? Who is responsible for this outcome?) mediate the effect of perceived emotion on cooperation.

Acknowledgments

This work was sponsored by the Fundação para a Ciência e a Tecnologia (FCT) grant #SFRH-BD-39590-2007; and, the Air Force Office of Scientific Research under grant FA9550-09-1-0507 and the National Science Foundation under grant IIS-0916858. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Axelrod R. (1984). *The Evolution of Cooperation*. New York, USA: Basic Books.

Beale, R., & Creed, C. (2009). Affective interaction: How emotional agents affect users. *Human-Computer Studies, 67*(2), 755-776. doi: 10.1016/j.ijhcs.2009.05.001

Bente, G., Feist, A., & Elder, S. (1996). Person Perception Effects of Computer-Simulated Male and Female Head Movement. *Journal of Nonverbal Behavior, 20*(4), 213–228. doi:10.1007/BF02248674

Bente, G., Kramer, N., Petersen, A., & de Ruiter, J. (2001). Computer Animated Movement and Person Perception: Methodological Advances in Nonverbal Behavior Research. *Journal of Nonverbal Behavior, 25*(3), 151-166. doi:10.1023/A:1010690525717

Berry, D., Butler, L., & De Rosis, F. (2005). Evaluating a realistic agent in an advice-giving task. *Journal of Human–Computer Studies*, *63*(3), 304–327. doi:10.1016/j.ijhcs.2005.03.006

Bixenstine, V., & Wilson, K. (1963). Effects of level of cooperative choice by the other player on choices in a prisoner's dilemma game, Part II. *Journal of Abnormal and Social Psychology, 67*(2), 139-147.

Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., & Bailenson, J. (2002). Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology. *Psychological Inquiry, 13*(2), 103-124. doi:10.1207/S15327965PLI1302_01

Boone, R., & Buck, R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior, 27*(3), 163–182. doi:10.1023/A:1025341931128

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *Journal of Human–Computer Studies 62*(2), 161–178. doi:10.1016/j.ijhcs.2004.11.002

Chapman, H., Kim, D., Susskind, J. & Anderson, A. (2009). In bad taste: Evidence for the oral origins of moral disgust. *Science, 323*(5918), 1222–1226. doi:10.1126/science.1165565

Creed, C., & Beale, R. (2008). Psychological responses to simulated displays of mismatched emotional expressions. *Interacting with Computers*, *20*(2), 225–239. doi:10.1016/j.intcom.2007.11.004

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, USA: G.P. Putnan's Sons.

Dawkins, R. (1976). *The selfish gene*. New York, USA: Oxford University Press.

De Cremer, D. (1999). Trust and fear of exploitation in a public goods dilemma. *Current Psychology, 18*(2), 153–163. doi: 10.1007/s12144-999-1024-0

de Melo, C., & Gratch, J. (2009). Expression of Emotions using Wrinkles, Blushing, Sweating and Tears. *Lecture Notes in Computer Science, Vol. 5773. Intelligent Virtual Agents* (pp. 188-200). Berlin, Germany: Springer-Verlag.

de Melo, C., Zheng, L., & Gratch, J. (2009). Expression of Moral Emotions in Cooperating Agents. In *Proceedings of Intelligent Virtual Agents 2009,* 301-307. doi:10.1007/978-3-642-04380-2_23

Dehn, D., & Van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies 52*(1), 1-22. doi:10.1006/ijhc.1999.0325

Deutsch, M., Epstein, Y., Canavan, D., & Gumpert, P. (1967). Strategies for inducing cooperation: An experimental study. *Journal of Conflict Resolution, 11*(3), 104-109. doi:10.1177/002200276701100309

Ellsworth, P., & Scherer, K. (2003). Appraisal Processes in Emotion. In: R. Davidson, K. Scherer, & H. Goldsmith, (Eds.), *Handbook of Affective Sciences* (pp. 572–595). New York, USA: Oxford University Press.

Frank, R. (1988). *Passions within reason: The strategic role of the emotions*. New York, USA: Norton.

Frank, R. (2004). Introducing moral emotions into models of rational choice. In A. Manstead, N. Frijda, & A. Fischer (Eds.), *Feelings and emotions* (pp. 422–440). New York, USA: Cambridge University Press.

Frijda, N., & Mesquita, B. (1994). The social roles and functions of emotions. In S. Kitayama & H. Markus (Eds.), *Emotion and culture: Empirical studies of mutual influence* (pp. 51–87). Washington, DC, USA: American Psychological Association.

Gong, L. (2007). Is happy better than sad even if they are both non-adaptive? Effects of emotional expressions of talking–head interface e-agents. *Journal of Human–Computer Studies 65*(3), 183–191. doi:10.1016/j.ijhcs.2006.09.005

Hamilton, W. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology, 7*(1), 17–52.

Hareli, S., & Hess, U. (2009). What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception. *Cognition & Emotion, 24*(1), 128-140. doi:10.1080/02699930802613828

Harford, T., & Solomon, L. (1967). ‘Reformed sinner’ and ‘lapsed saint’ strategies in the prisoner’s dilemma game. *Journal of Conflict Resolution, 11*(1), 345-360. doi:10.1177/002200276701100109

Helson, H. (1964). *Adaptation-level theory*. New York, USA: Harper & Row.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(3), 383-451.

Hilty, J., & Carnevale, P. (1993) Black-Hat/White-Hat Strategy in Bilateral Negotiation. *Organizational Behavior and Human Decision Processes, 55*(3), 444-469. doi:10.1006/obhd.1993.1039

Hone, K. (2006). Empathic agents to reduce user frustration: the effects of varying agent characteristics. *Interacting with Computers 18*(2), 227–245. doi:10.1016/j.intcom.2005.05.003

Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition and Emotion*, *13*(5), 505–521. doi:10.1080/026999399379267

Keltner, D., & Kring, A. (1998). Emotion, Social Function, and Psychopathology. *Review of General Psychology*, *2*(3), 320–342. doi:10.1037//1089-2680.2.3.320

Kiesler, S., Waters, K., & Sproull, L. (1996). A Prisoner's Dilemma Experiment on Cooperation with Human-Like Computers. *Journal of Personality and Social Psychology, 70*(1), 47–65.

Klein, J., Moon, Y., & Picard, R. (2002). This computer responds to user frustration: theory, design, and results. *Interacting with Computers 14*(2), 119–140. doi:10.1016/S0953-5438(01)00053-4

Kollock, P. (1998). Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology, 24,* 183-214. doi: 10.1146/annurev.soc.24.1.183

Krumhuber, E., Manstead, A. & Kappas, A. (2007). Facial Dynamics as Indicators of Trustworthiness and Cooperative Behavior. *Emotion*, *7*(4), 730–735. doi:10.1037/1528-3542.7.4.730

Lester, J., Converse, S., Kahler, S., Barlow, T., Stone, B., & Bhogal, R. (1997). The persona effect: affective impact of animated pedagogical agents. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 359–366). New York, USA: ACM Press.

Lewis, M. (2008). Self-Conscious Emotions: Embarrassment, Pride, Shame, and Guilt. In L. Michael & J. Haviland-Jones (Eds.) *Handbook of Emotions* (pp. 623–636). New York, USA: The Guilford Press.

Lim, Y., & Aylett, R. (2007). Feel the difference: a guide with attitude!. *Lecture Notes in Computer Science, Vol. 4722. Intelligent Virtual Agents* (pp. 317–330). Berlin, Germany: Springer-Verlag.

Liu, K., & Picard, R. (2005). Embedded empathy in continuous, interactive health assessment. Paper presented at the Computer–Human Interaction Workshop on Computer–Human Interaction Challenges in Health Assessment.

Maldonado, H., Lee, J., Brave, S., Nass, C., Nakajima, H., Yamada, R., Iwamura, K., & Morishima, Y. (2005). We learn better together: enhancing eLearning with emotional characters. In T. Koschmann, D. Suthers, & T. Chan (Eds.), *Computer Supported Collaborative Learning 2005: The Next 10 Years!* (pp. 408–417). New Jersey, USA: Lawrence Erlbaum Associates.

Marsella, S., Gratch, J., & Petta, P. (2010). Computational Models of Emotion. In K. Scherer, T. Bänziger, & E. Roesch (Eds.). *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing* (pp. 21-45). New York, USA: Oxford University Press.

Mehu, M., Grammer, K., & Dunbar, R. I. (2007). Smiles when sharing. *Evolution and Human Behavior, 28*(6), 415–422. doi:10.1016/j.evolhumbehav.2007.05.010

Morris, M., & Keltner, D. (2000). How emotions work: An analysis of the social functions of emotional expression in negotiations. *Research in Organizational Behavior, 22*, 1–50. doi:10.1016/S0191-3085(00)22002-9

Nass, C., Steuer, J., & Tauber, E. (1994). Computers are Social Actors. *Proceedings of the SIGCHI conference on Human factors in computing* (pp. 72-78). New York, USA: ACM Press.

Ortony A, Clore G, & Collins A. (1988). *The Cognitive Structure of Emotions*. New York, USA: Cambridge University Press.

Prendinger, H., Mayer, S., Mori, J., & Ishizuka, M. (2003). Persona effect revisited. Using bio-signals to measure and reflect the impact of character-based interfaces. *Proceedings* *of the Fourth International Working Conference On Intelligent Virtual Agents* (pp. 283–291). Berlin, Germany: Springer-Verlag.

Pruitt, D., & Carnevale, P. (1993). *Negotiation in social conflict*. Pacific Grove, CA, USA: Brooks/Cole.

Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York, USA: Cambridge University Press.

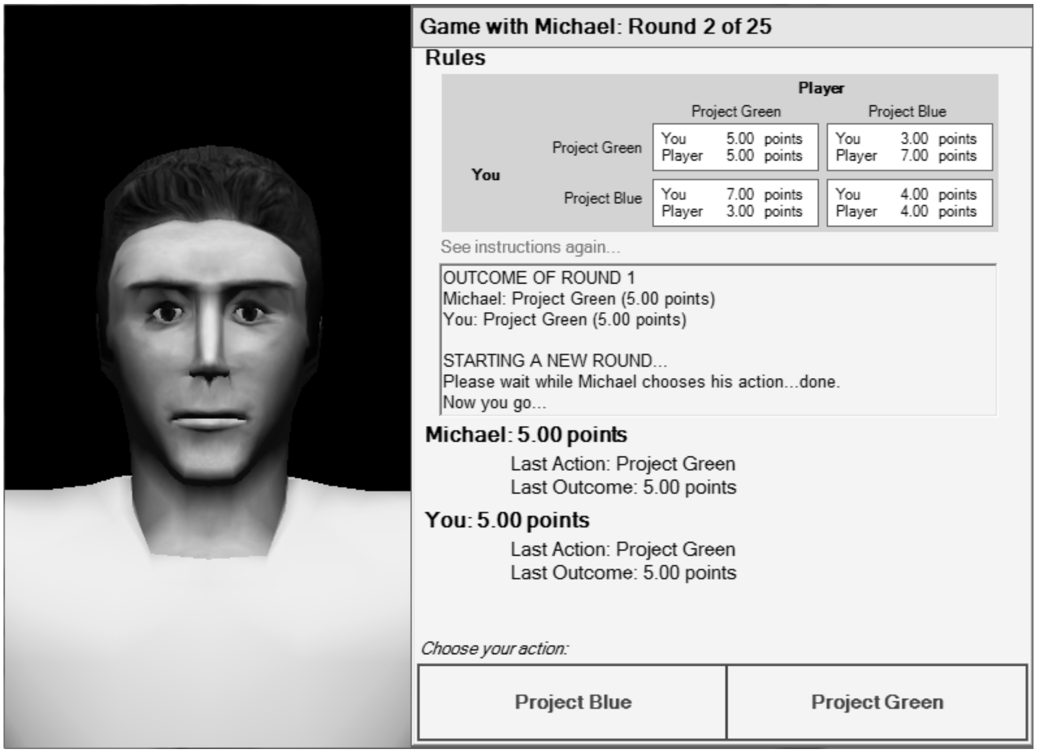
Scharlemann, J., Eckel, C., Kacelnik, A., & Wilson, R. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology, 22*(5), 617–640. doi:10.1016/S0167-4870(01)00059-9

Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T. & Bonnet, K. (2010). Emotional expressivity as a signal of cooperation. *Evolution and Human Behavior, 31*(2), 87–94. doi:10.1016/j.evolhumbehav.2009.09.006

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*(1), 35–57.

Van Mulken, S., André, E., & Muller, J. (1998). The persona effect: how substantial is it? *Proceedings of HCI on People and Computers XIII* (pp. 53–66). London, UK: Springer-Verlag.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598. doi:10.1111/j.1467-9280.2006.01750.x



Game with Michael: Round 2 of 25

		Player: Project Green	Player: Project Blue
You	Project Green	You 5.00 points Player 5.00 points	You 3.00 points Player 7.00 points
	Project Blue	You 7.00 points Player 3.00 points	You 4.00 points Player 4.00 points

```
OUTCOME OF ROUND 1
Michael: Project Green (5.00 points)
You: Project Green (5.00 points)

STARTING A NEW ROUND...
Please wait while Michael chooses his action...done.
Now you go...
```

Michael: 5.00 points
Last Action: Project Green
Last Outcome: 5.00 points

You: 5.00 points
Last Action: Project Green
Last Outcome: 5.00 points

Figure 1. The software used in the experiment. During game play, the payoff matrix is shown on the top right, the outcome of the previous round in the upper mid right, the total outcome and the actions in the previous round in the lower mid right, the possible actions on the bottom right and the real-time animation of the agent on the left.

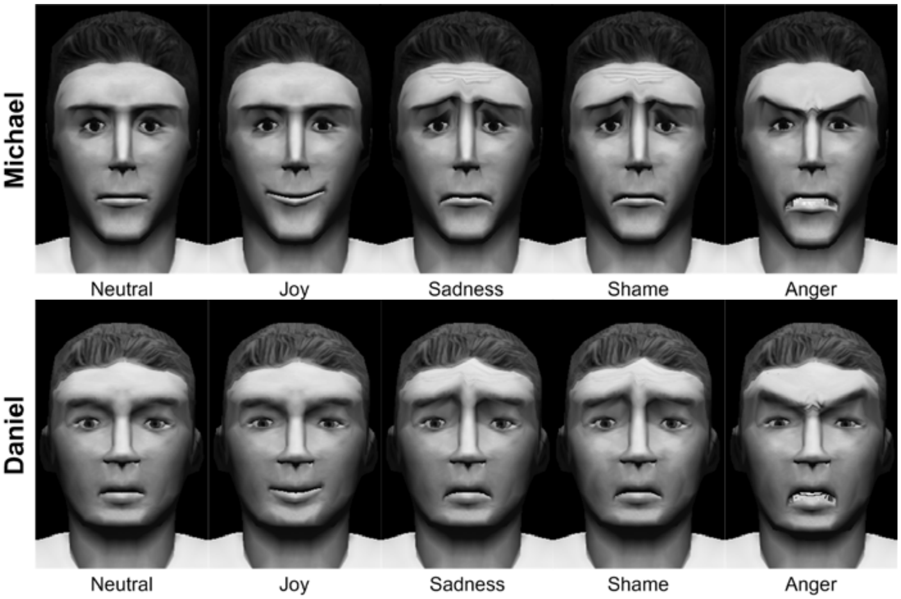


Figure 2. The agent bodies - Michael and Daniel - and their facial displays. Shame is distinguished from sadness by blushing of the cheeks.

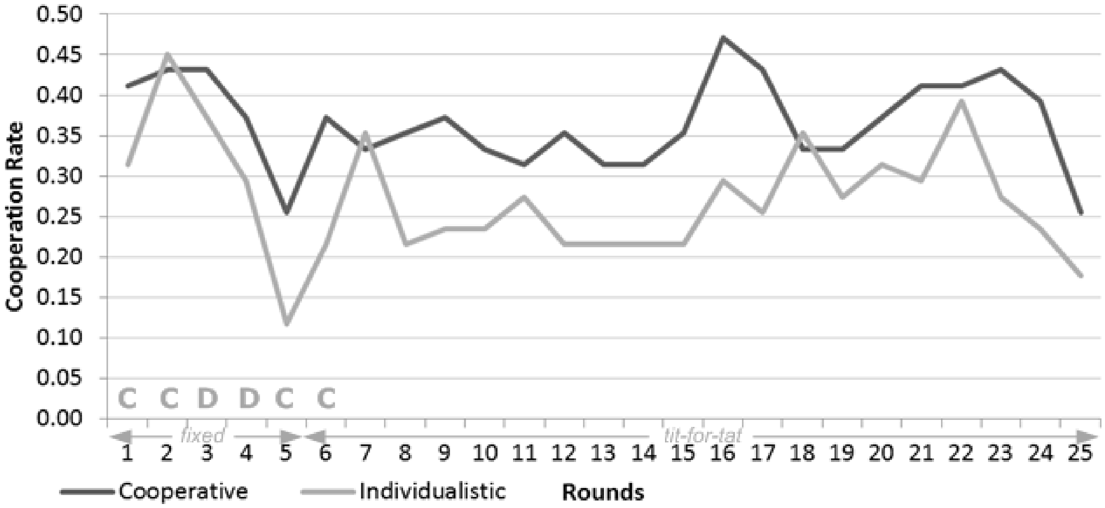


Figure 3. Evolution of cooperation rate across rounds. The agent strategy is marked above the horizontal axis: ‘C’ stands for cooperation and ‘D’ for defection.