People's Biased Decisions to Trust and Cooperate with Agents that Express Emotions

Celso M. de Melo[1], Peter J. Carnevale[1], and Jonathan Gratch[2]

[1]USC Marshall School of Business, Los Angeles, CA 90089-0808, USA
demelo@usc.edu, peter.carnevale@marshall.usc.edu
[2]Institute for Creative Technologies, University of Southern California, 12015 Waterfront Drive, Building #4 Playa Vista, CA 90094-2536, USA
gratch@ict.usc.edu

Abstract. Research in the behavioral sciences shows that emotion expressions impact people's decisions to trust and cooperate with others in situations where self and collective interests collide. Building on such findings, computer scientists have shown that emotion expressions in agents can also impact people's decision making. However, recent findings in neuroeconomics reveal that people systematically show different behavior and brain activation patterns in decision making tasks with computers, when compared to humans. These findings suggest a bias people might have with respect to autonomous agents and, in particular, agents that express emotions. To clarify this, the paper presents a novel experiment where participants engaged in the iterated prisoner's dilemma, for clear financial stakes, with counterparts, either agents or humans, that showed facial displays of emotion that were compatible with a cooperative (e.g., smile after mutual cooperation) or competitive (e.g., smile after exploiting the participant) goal orientation. The results showed that participants cooperated, as expected, more with cooperative than competitive counterparts but, also revealed that people trusted and cooperated more with a human that showed cooperative displays than an agent that showed the exact same displays. We discuss implications of such a bias for trust and cooperation in human-agent interaction.

Keywords. Emotion, Decision Making, Trust, Cooperation, Bias

1 Introduction

In many social settings one has to decide between following short-term self-interest or taking a risk and trusting that others will reciprocate cooperation [1]. In such social dilemmas people look for cues that others are trustworthy and willing to cooperate [2]. One such cue is non-verbal behavior, in particular displays of emotion [3-5]. Indeed, several researchers have now acknowledged that emotions help regulate social interaction and serve important social functions such as communicating one's beliefs, desires and intentions to others [6-8]. Supporting this view, empirical evidence in the behavioral sciences shows that people's emotion expressions can impact our judgments of their trustworthiness and our decisions to trust them [9-12]. Complementarily, computer scientists have begun exploring how expressions of emotions in autonomous agents impact people's decision making (e.g., [13-17]). In a particularly relevant experiment to this paper [16], we showed that, in the iterated prisoner's dilemma, people would trust and cooperate more with agents that showed, in their virtual faces, cooperative emotion displays (e.g., smile after mutual cooperation) than agents that showed competitive emotion displays (e.g., smile after exploiting the participant). It would appear, then, that findings from the human-human interaction literature about the impact of emotion expressions on people's decision to trust and cooperate could easily propagate to human-agent interaction settings. However, recent findings in the emerging field of neuroeconomics cast doubt on this conclusion.

Recent findings show that people systematically reach different decisions and show different patterns of brain activation with computers in the exact same decision making tasks, for the exact same financial incentives, when compared to humans. Gallagher et al. [18] showed that when people played the rock-paper-scissors game with a human there was activation of the medial prefrontal cortex (MPFC), which had previously been shown to be involved in mentalizing (i.e., inferring of the other's beliefs, desires and intentions); however, no such activation occurred when people engaged with a computer that followed a predefined algorithm to make the choice. McCabe et al. [19] found a similar pattern when people played the trust game with humans in comparison to a probabilistic algorithm; Riedl et al. [20] further replicated this result with embodied agents, i.e., agents with three-dimensional bodies and faces. In the prisoner's dilemma, Rilling et al. [21] and Krach et al. [22] showed that people tended to cooperate more with humans than computers and, once again, brain regions associated with mentalizing such as the MPFC, the rostral anterior cingulate cortex and the right temporo-parietal junction, were only activated with humans; in contrast, Kircher et al. [23] showed no difference in cooperation rates between humans and computers, despite reporting the usual increased brain activity with humans. In an influential paper, Sanfey et al. [24] showed that people were more willing to accept unfair offers in the ultimatum game from a computer than from a human. Moreover, their results revealed that the bilateral anterior insula–a region usually associated with the experience of negative emotions–showed higher activation when people received unfair offers from humans than from computers, thus suggesting that increased negative emotion explained the discrepancies in decision making behavior. Complementing this work, van't Wout et al. [25] showed that unfair offers in the ultimatum game led skin conductance–an autonomic index of affective state–to raise with humans but not with computers. In sum, these findings suggest people show a systematic bias with computers in the way they reach decisions to trust or cooperate in social decision making tasks when compared to humans.

The goal of this paper is, thus, to understand whether people's decisions to trust and cooperate would also show a bias when engaged in social decision making with emotional agents when compared to emotional humans. To accomplish this we present a novel experiment where participants engaged in the iterated prisoner's dilemma with an emotional counterpart that was either an embodied agent or a virtual representation of a human (i.e., an avatar). Counterparts showed emotion displays that were either compatible with a cooperative or competitive goal orientation. Our main measure was cooperation rate over all rounds. Following our previous findings [15-17], one could expect people to cooperate more with cooperative than competitive counterparts; however, what is not clear is whether there will be a quantitative or qualitative difference in the effects when we compare interactions with agents vs. avatars.

2 Method

In choosing the prisoner's dilemma we follow several other researchers that used social dilemmas to operationalize the study of trust (e.g., [26, 27]). Social dilemmas are ideal because they emphasize expectations of others' benevolence in situations that involve conflict between selfish and collective interests. In this experiment, participants played 20 rounds. The payoff matrix we used is shown in Table 1. Following the approach by Kiesler, Waters and Sproull [28], the prisoner's dilemma was also recast as an investment game.

Table 1. Payoff matrix for the prisoner's dilemma.

		Counterpart	
		Cooperation	Defection
Participant	Cooperation	Counterpart: 6 pts Participant: 6 pts	Counterpart: 10 pts Participant: 0 pts
	Defection	Counterpart: 0 pts Participant: 10 pts	Counterpart: 3 pts Participant: 3 pts

Participants engaged with either an autonomous agent or an avatar. Autonomous agents were always referred to as "computer agents" and were described to the participants as "a computer program that was designed to make decisions just like other people". Avatars were described as "the players' visual representation in the game". Participants were asked to choose an avatar for themselves, of the same gender, and were informed that their avatar "would be visible to the other player" and that the they "would be able to control aspects of the avatar's behavior which would be visible to the other player, and vice-versa". In reality, participants always played with a computer program that followed the same strategy: tit-for-tat, starting with a defection. To make this deception believable, we implemented a server that matched pairs of participants that were supposed to engage with other participants; participants would then proceed in lockstep throughout the task but the responses they'd see always followed the tit-for-tat strategy. Participants were also made to believe they were engaging with a participant of the same gender when, in fact, this might have not been the case. Lastly, participants were told that the identities of other participants would be concealed and the software always referred to the human counterpart as "anonymous".

Participants were matched with a counterpart that either showed cooperative or competitive facial displays. Following our previous findings [15-17], we defined the *expressively cooperative* counterpart (Table 2, top), which displays joy in mutual cooperation and guilt when it exploits the participant, and the *expressively competitive* counterpart (Table 2, bottom), which displays joy when it exploits the participant and guilt in mutual cooperation. The rationale for the cooperative agent is that joy after mutual cooperation signals an intention to cooperate, whereas guilt after exploitation signals regret and an apology for the transgression; the rationale for the competitive agent is that joy after exploitation signals an intention to compete, whereas guilt after mutual cooperation signals regret for missing the chance to exploit the participant. Given our previous evidence [15-17], we expected people to cooperate more with cooperative than competitive counterparts. However, in these studies, some ambiguity was left regarding whether the counterpart was an agent or a person (e.g., agents were referred by a name, such as "Ethan"). Thus, the previous results do not clarify whether people react differently to emotions displayed by an agent in comparison to those of a human.

Table 2. Facial displays for the emotional counterparts.

Expressively Cooperative		*Counterpart*	
		Cooperation	Defection
Participant	Cooperation	Joy	Guilt
	Defection	Neutral	Neutral

Expressively Competitive		*Counterpart*	
		Cooperation	Defection
Participant	Cooperation	Guilt	Joy
	Defection	Neutral	Neutral

2.1 Design

The experiment followed a 2 × 2 factorial design: *Emotion Displays* (Cooperative vs. Competitive) × *Agency* (Agent vs. Avatar). We used the same emotion facial displays that were validated and used in our previous experiments [15-17]: joy is expressed through a smile and contraction of the corrugator supercilii (eyes); guilt is expressed through lowering of the zygomaticus, blushing, bowing of the head and aversion of gaze. Male and female avatars used in the experiment are shown in Figure 1.

2.2 Measures

Our main dependent variable was cooperation rate, i.e., the number of times participants cooperated over all rounds. After completing the task, we also asked about perception of the counterpart's trustworthiness (scale went from 1, *not at all*, to 7, *very much*): How trustworthy was the other party? Lastly, to validate that participants were correctly perceiving some counterparts as agents and others as humans we asked them, before concluding the experiment, to rate the counterpart according to the following pairs of adjectives on a 7-point scale (e.g., for Fake-Natural, 1 corresponded to Fake and 7 to Natural): Robot like-Human like; Fake-Natural; Unconscious-Conscious; Artificial-Lifelike; Stagnant-Lively; Mechanical-Organic; Inert-

Interactive; Apathetic-Responsive; and, Computer-Human. The selection of these adjectives was based on existent scales pertaining to anthropomorphism [29], the "uncanny valley" effect (e.g. [30]) and the experience of social presence in virtual environments (e.g., [31]).

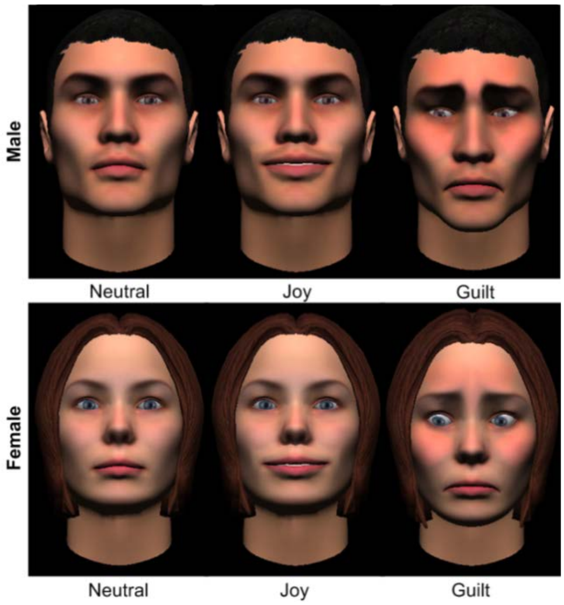

Fig. 1. The emotion facial displays used in the experiment.

2.3 Participants

One-hundred and twenty six participants were recruited at USC's Marshall School of Business. This resulted in approximately 30 participants per condition. Regarding gender, 69.7% were males. Age distribution was as follows: 21 years and under, 70.6%; 22 to 34 years, 29.4%. Most participants were undergraduate students (95.8%) majoring in Business related courses and with citizenship from the United States (81.5%). The incentive to participate followed standard practice in experimental economics [32]: first, participants were given school credit for their participation; second, with respect to their goal in the task, and similarly to our previous experiments [15-17], participants were instructed to earn as many points as possible as the total amount of points would increase their chances of winning a lottery of $100. Upon completion of the experiment participants were verbally debriefed about the deception pertaining to the avatar conditions.

3 Results

Participants that did not experience both joy and guilt with the counterpart[1]–i.e., our experimental manipulation–were excluded from analysis (though keeping them would lead to the same pattern of results). After exclusion, 84 participants remained for analysis.

3.1 Manipulation Check

The nine adjective classification questions were highly correlated (Cronbach α = .972) and, thus, were averaged into a single measure we called anthropomorphism. We then ran an Emotion Displays × Agency ANOVA which revealed no main effect of Emotion Displays, $F(1, 80) = 1.13$, $p = .291$, but, as expected, confirmed a main effect of Agency, $F(1, 80) = 4.48$, $p = .037$, partial $\eta^2 = .053$: people perceived the human counterparts ($M = 4.87$, $SD = 1.54$) to be more anthropomorphic than the agent counterparts ($M = 4.12$, $SD = 1.64$). The Emotion Displays × Agency interaction was not significant, $F(1, 80) = .541$, $p = .464$.

3.2 Cooperation Rate

The means and standard errors for cooperation rate are shown in Figure 2. We ran an Emotion Displays × Agency ANOVA to analyze this variable. The results revealed no main effect of Agency, $F(1, 80) = .612$, $p = .436$, and no Emotion Displays × Agency interaction, $F(1, 80) = .490$, $p = .486$. However, there was a main effect of Emotion Displays, $F(1, 80) = 5.68$, $p = .020$, partial $\eta^2 = .066$: people cooperated more with cooperative counterparts ($M = .69$, $SD = .27$) than competitive counterparts ($M = .55$, $SD = .28$). To further tease this effect apart, we split the data across Agency and ran independent *t* tests to compare cooperation rates between cooperative and competitive counterparts. This analysis revealed that, for agents, people cooperated more with cooperative ($M = .64$, $SD = .26$) than competitive agents ($M = .54$, $SD = .30$) but this result did not reach significance, $t(37) = 1.12$, $p = .269$, $r = .181$. For humans, people cooperated more with cooperative ($M = .73$, $SD = .26$) than competitive humans ($M = .55$, $SD = .28$) and this result was significant, $t(43) = 2.31$, $p = .026$, $r = .332$.

[1] Notice this paradigm did not guarantee participants would experience all outcomes in the prisoner's dilemma task.

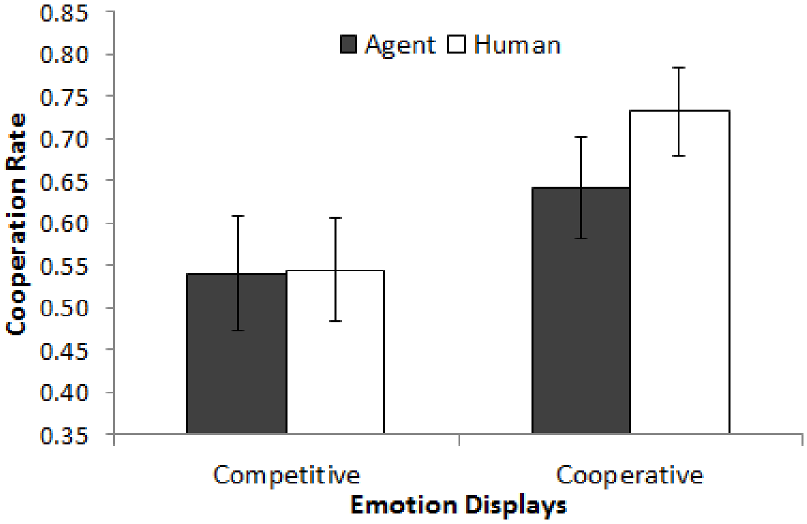

Fig. 2. Means (and standard errors) for cooperation rate.

3.3 Perceptions of Trustworthiness

We ran an Emotion Displays × Agency ANOVA to analyze perceptions of trustworthiness. The results revealed a trend for a main effect of Agency, $F(1, 80) = 5.45$, $p = .147$, partial $\eta^2 = .026$: people tended to perceive agents ($M = 4.87$, $SD = 1.77$) as less trustworthy than humans ($M = 5.42$, $SD = 1.47$). There was also a stronger trend for a main effect of Emotion Displays, $F(1, 80) = 9.54$, $p = .056$, partial $\eta^2 = .045$: people tended to perceive cooperative counterparts ($M = 5.50$, $SD = 1.50$) as more trustworthy than competitive counterparts ($M = 4.80$, $SD = 1.70$). Finally, there was no Agency x Emotion Displays interaction, $F(1, 80) = .733$, $p = .593$.

4 Discussion

In this paper we had people engage in a social dilemma, with clear financial consequences, with either emotional autonomous agents or emotional humans. The results confirmed that people reach different decisions about trust and cooperation when engaged with agents that honestly portray themselves as computers compared to if they portray themselves as human. Merely this belief – even though the financial incentives and the agent's appearance, decisions, and expressions were identical – had a powerful effect. In either case, emotion displays appeared to shape participants' willingness to cooperate, but these effects were only significant, and the effect sizes much larger ($r = .332$ vs. $r = .181$) when playing against a presumed human opponent. Strikingly, as cooperative emotions promoted greater cooperation rates and thus greater individual rewards, participants were able to earn more money when they were deceived about the true nature of the agent. The results, thus, confirm a bias people display towards agents, even if these agents are described to be "designed to make decisions just like other people".

These results are compatible with recent findings in the emerging neuroeconomics field that suggest that brain regions usually associated with mentalizing tend to show higher activation patterns in decision making tasks when people believe they are engaging with humans rather than computers. Effectively, we have argued elsewhere [15, 17] that a key for the social effects of emotion expressions is the information people retrieve from such displays about the other's beliefs, desires and intentions. In this sense, a higher activation of the mentalizing brain regions with humans might have meant people tried harder to infer the human's mental states from their emotion displays, which then led to increased effects when compared to agents. Sanfey et al.'s [24] proposal that people experience higher activation of brain regions associated to emotion with humans rather than computers is also compatible with our findings. In our case, it is plausible people might have felt, for instance, higher positive emotion upon seeing a smile after mutual cooperation coming from a human than from an agent and that, in turn, led to increased trust and cooperation with the former.

Our results are also in line with findings in the human-computer interaction and communication literatures that, despite not focusing on people's decision making in standard experimental economics tasks, suggest people experience higher social presence [33, 34], inhibition [35], learning [36], flow [37], arousal [38] and engagement [34] with humans than agents. From a theoretical point of view, our results contrast with the "computer as social actors" theory [39, 40] that argues people treat machines (such as agents) that display social behavior in the same manner as humans. Instead, our results are more in line with Blascovich and colleagues' social influence theory [41, 42] that suggests people are more likely to be influenced by a virtual entity (such as an agent) the more this entity is believed to be controlled by a human (which they refer to as "agency"). The theory further suggests that to achieve social influence one could compensate for an autonomous agent's lack of agency by increasing the agent's (behavioral and visual) realism. This is, in fact, a promising line of future inquiry.

The results for perceptions of trustworthiness were more subtle. Despite suggesting people find humans more trustworthy than agents, the results did not reach significance. This might suggest that, despite showing clear evidence for a bias in their cooperation behavior, people are more reluctant to acknowledge (i.e., self-report) such a bias. An alternative explanation is that the bias is occurring unconsciously and, therefore, self-report measures are unable to fully capture it.

The question that remains to be answered, of course, is: Why are people reaching different decisions about trust and cooperation with agents when compared to humans in the same social decision making situations? The goal of the paper was not to answer this question but, to show that there is a bias and call the autonomous agents community's attention to it. Perhaps, this bias only reflects people's current suspicion about the ability of a machine to "have a mind", i.e., a mind that is worthy of mentalizing as is the mind of a human. If this is the case, then the research agenda should focus on, incrementally, simulating the cognitive abilities we see in humans (e.g., learning) and demonstrating they exist to the person interacting with our agents until we cross the threshold above which people don't distinguish between humans and agents. Or, perhaps, this bias reflects an inherent difference between humans and machines that cannot be surpassed by adding more "intelligence" to machines. Perhaps people will always treat agents as out-group members and expose them to corresponding prejudice [43]. If this is the case, then the research agenda should focus, instead, on understanding (and accepting) the similarities and differences between human-human and agent-human social decision making.

Acknowledgments. This research was supported in part by the Air Force Office of Scientific Research (AFOSR) and grant 1211064 from the NSF. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Dawes, R.: Social dilemmas. Annu Rev Psychol 31, 169-193 (1980)
2. Kollock, P.: Social dilemmas: The anatomy of cooperation. Annu Rev Sociol 24, 183-214 (1998)
3. Frank, R.: Introducing moral emotions into models of rational choice. In: A. S. R. Manstead, N. Frijda & A. Fischer (Eds.), Feelings and emotions, pp. 422-440. New York, NY: Cambridge University Press (2004)
4. Trivers, R.: The evolution of reciprocal altruism. Q Rev Biol 46, 35-57 (1971)
5. Boone, R., Buck, R.: Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. J Nonverbal Behav 27, 163-182 (2003)
6. Frijda, N., Mesquita, B.: 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: American Psychological Association (1994)
7. Keltner, D., Haidt, J.: Social functions of emotions at four levels of analysis. Cognition and Emotion 13, 505–521 (1999)
8. Keltner, D., Kring, A.: Emotion, social function, and psychopathology. Rev Gen Psychol 2(3), 320–342 (1998)
9. Scharlemann, J., Eckel, C., Kacelnik, A., Wilson, R.: The value of a smile: Game theory with a human face. J Econ Psychol 22, 617-640 (2001)
10. Krumhuber, E., Manstead, A. Kappas, A.: Facial dynamics as indicators of trustworthiness and cooperative behavior. Emotion 7, 730-735 (2007)
11. Campellone, T., Kring, A.: Who do you trust? The impact of facial emotion and behaviour on decision making. Cognition and Emotion, 1-18 (2012)
12. Van Kleef, G., De Dreu, C., Manstead, A.: An interpersonal approach to emotion in social decision making: The emotions as social information model. Ad Exp Soc Psychol 42, 45-96 (2010)
13. Beale, R., & Creed, C.: Affective interaction: How emotional agents affect users. Int J Hum-Comput St 67, 755-776 (2009)
14. de Melo, C., Carnevale, P., Gratch, J.: The effect of expression of anger and happiness in computer agents on negotiations with humans. In: Proceedings of Autonomous Agents and Multiagent Systems (2011)
15. de Melo, C.: The interpersonal effect of emotion in decision-making and social dilemmas. Ph.D. diss., Department of Computer Science, University of Southern California, Los Angeles, CA (2012)

16. de Melo, C., Carnevale, P., Gratch, J.: The impact of emotion displays in embodied agents on emergence of cooperation with people. Presence 20, 449-465 (2012)
17. de Melo, C., Carnevale, P., Read, S., Gratch, J.: Reverse appraisal: The importance of appraisals for the effect of emotion displays on people's decision-making in a social dilemma. In: Proceedings of 34th Annual Meeting of the Cognitive Science Society (2012)
18. Gallagher, H. et al.: Imaging the intentional stance in a competitive game. NeuroImage 16, 814-821 (2002)
19. McCabe, K. et al.: A functional imaging study of cooperation in two-person reciprocal exchange. PNAS 98, 11832-11835 (2001)
20. Riedl, R. et al.: Trusting humans and avatars: Behavioral and neural evidence. In: Proceedings of the 32nd Int. Conference on Information Systems (2011)
21. Rilling, J. et al.: A neural basis for social cooperation. Neuron 35, 395-405 (2002)
22. Krach, S. et al.: Can machines think? Interaction and perspective taking with robots investigated via fMRI. PLoS ONE 3, 1-11 (2008)
23. Kircher, T. et al.: Online mentalising investigated with functional MRI. Neurosci Lett 454, 176-181 (2009)
24. Sanfey, A. et al.: The neural basis of economic decision-making in the ultimatum game. Science 300, 1755-1758 (2003)
25. van't Wout. M. et al.: Affective state and decision-making in the ultimatum game. Exp Brain Res 169, 564-568 (2006)
26. Yamagishi, T.: Trust: The evolutionary game of mind and society. New York, NY: Springer (2011)
27. Balliet, D., Van Lange, P.: Trust, conflict, and cooperation: A meta-analysis. Psychol Bull, 1-23 (2012)
28. Kiesler, S., Waters, K., Sproull, L.: A prisoner’s dilemma experiment on cooperation with human-like computers. J Pers Soc Psychol 70, 47-65 (1996)
29. Epley, N., Waytz, A., Cacioppo, J.: On seeing human: A three-factor theory of anthropomorphism. Psych Rev 114, 864-886 (2007)
30. Bartneck, C., Kulic, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Rob 1, 71-81 (2009)
31. Harms, C., Biocca, F.: Internal consistency and reliability of the networked minds measure of social presence. In: Proceedings of 7th Annual Workshop: Presence (2004)
32. Hertwig, R. and Ortmann, A.: Experimental practices in economics: A methodological challenge for psychologists? Behav Brain Sci 24, 383-451 (2001)
33. Gajadhar, B., de Kort, Y., Ijsselsteijn, W.: Shared fun is doubled fun: Player enjoyment as a function of social setting. In: Proceedings of the Fun and Games Conference (2008)
34. Ravaja, N.: The psychophysiology of digital gaming: The effect of a non co-located opponent. Media Psych 12, 268-294 (2009)
35. Hoyt, C., Blascovich, J. and Swinth, K.: Social inhibition in immersive virtual environments. Presence 12, 183-195 (2003)
36. Okita, S., Bailenson, J., Schwartz, D.: The mere belief of social interaction improves learning. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2007)
37. Weibel, D., Wissmath, B., Habegger, S., Steiner, Y., Groner, R.: Playing online games against computer- vs. human-controlled opponents: Effects on presence, flow, and enjoyment. Compu Hum Behav 24, 2274-2291 (2008)
38. Katsyri, J., Hari, R., Ravaja, N., Nummenmaa, L.: The opponent matters: Elevated fMRI reward responses to winning against a human versus a computer opponent during interactive video game playing. Cerebral Cortex, Sep/2012 (2012)

39. Nass, C., Steuer, J., Tauber, E.: Computers are social actors. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (1994)
40. Reeves, B., Nass, C.: The media equation: How people treat computers, television, and new media like real people and places. New York, NY: Cambridge University Press (1996)
41. Blascovich, J., McCall, C.: Social influence in virtual environments. In: K. Dill (Ed.), The Oxford Handbook of Media Psychology, pp.305-315. New York, NY: Oxford University Press (2013)
42. Blascovich, J. et al.: Immersive virtual environment technology as a methodological tool for social psychology. Psychol Inq 13, 103-124 (2002)
43. Brewer, M. The psychology of prejudice: Ingroup love and outgroup hate? J Soc Issues 55, 429-444 (1999)