# The effect of embodiment and competence on trust and cooperation in human–agent interaction

Philipp Kulms, Stefan Kopp

Social Cognitive Systems Group, Center of Excellence 'Cognitive Interaction Technology' (CITEC), Faculty of Technology, Bielefeld University, Germany
{pkulms, skopp}@techfak.uni-bielefeld.de

**Abstract.** Success in extended human–agent interaction depends on the ability of the agent to cooperate over repeated tasks. Yet, it is not clear how cooperation and trust change over the course of such interactions, and how this is interlinked with the developing perception of competence of the agent or its social appearance. We report findings from a human–agent experiment designed to measure trust in task-oriented cooperation with agents that vary in competence and embodiment. Results in terms of behavioral and subjective measures demonstrate an initial effect of embodiment, changing over time to a relatively higher importance of agent competence.

**Keywords:** Human–agent interaction; trust; cooperation; embodiment

## 1 Introduction

It has often been noted that future interaction with technology may be designed like a cooperation between partners with complementary competencies [5]. In such teams, each agent has some degree of autonomy to handle dynamic situations and to make decisions within uncertain situations. As part of the cooperation, agents may plan and suggest to their partners – human or artificial – possible actions. Disentangling under which conditions humans accept such approaches and benefit from them is crucial. One key aspect is that users are willing and able to trust an agent. However, it is unclear how user trust is related to the perceived capabilities of an agent, in particular regarding the altering abilities of learning agents with possibly unanticipated behavior. Shaping the social interaction with such agents may be a key variable in those situations. Interactions with virtual agents are known to elicit social effects similar to human–human interaction [17]. In this paper we present work that investigates the potential of IVAs to support trust and how it develops in, and influence an ongoing human–agent cooperation. In particular, we present a human–agent cooperation study to investigate how perceived competence and the visual presence of a virtual agent affect the interaction and user trust.

## 2 Theoretical Background

**Social Factors of Trust and Cooperation.** Trust is the willingness of an agent (trustor) to be vulnerable to the actions of another agent (trustee) based on the expectation that the trustee will perform a particular action [18]. The trustee has characteristics that help the trustor to decide whether placing its trust in this agent is risky or not. These characteristics (ability, benevolence, and integrity) form the trustee's trustworthiness and promote trust, but trust and trustworthiness are not identical [18]. Ability (i.e., competence) as a "can-do" component describes the extent to which the trustee can enact its motives toward a specific goal, while benevolence and integrity as "will-do" descriptions pertain to whether the trustee wants to use its abilities to act in the best interest of the trustor [8]. Trust and cooperation are often used interchangeably [12]. Indeed, trust facilitates cooperation and vice versa [9], yet cooperation is also possible without trust. Another misconception is the assumed equality behind trust in and credibility of computers. In contrast to trust, credibility is a perceived quality and phrases like "trusting in information" or "believing the output" refer to credibility, not trust [14]. In cooperative relations the goals of two agents are positively related, that is, if one agent increases the chances of achieving its goal, the other agents chances to achieve its goal are also promoted [12]. Various reasons can lead to positive goal interdependence: a necessary division of work to achieve otherwise unattainable task goals, reward structures based on joint achievements, sharing of resources, being faced with the same obstacle, or holding common membership of a social group [11].

**Cooperation in Human–Computer Interaction.** Computers are increasingly understood as partners that people affiliate with [21]. Recent evidence indicates that humans are able to work together with computers in complex task settings (see [5] for an overview). This evolution requires that computers and machines accurately communicate their trustworthiness, even if they are competent, and that humans develop an appropriate level of trust that matches the trustworthiness of the output, e.g. decision support [20]. Accordingly, Dautenhahn [10] referred to socially intelligent agents as agents that offer or mediate cooperation and problem solving through social abilities similar to humans. Examples of human–agent settings explain the importance of social attributions: participants responded with commitment to agent-led teams, yet agents gained less trust and fairness than humans [25]. Other work has shown that agent trustworthiness correlates with both agent and team performance [15]. Human-like interfaces offer unique possibilities to design meaningful task-oriented interactions through nonverbal communication [7] or multimodal cues [24]. Virtual agents are an instantiation of such human-like interfaces. People interacting with virtual agents usually report a more personal experience, a phenomenon tracing back to humans' tendency to mindlessly applying social rules to computers [22]. Virtual agents operating as decision support yield social benefits that go beyond common technology adoption explanations [23]. Through their visual presence, they provide support and persuasion toward a desirable outcome (see [2] for an overview). In a study that investigated the effect of the human-like interface in a dialog-based setting, a virtual agent elicited stronger social responses than text-based interaction [1]. On the flip side, presenting virtual agents along with tasks may incur costs on memory performance [3] and heightens expectations in realism, leading to decreased willingness to cooperate with the agent [16] (see also [4]). The context in which virtual agents are embedded plays an important role. Early work emphasized their role as personal assistants that mediate between the user and the interaction goals with applications ranging from pedagogical agents to information systems, sales agents, and museum guides to name a few examples. In the past few years, virtual agents have been increasingly considered as a source of simple yet meaningful cues that affect perceived trustworthiness and people's willingness to cooperate with them in social dilemmas (e.g. [19]). This perspective holds that people do not blindly guess their virtual counterpart's upcoming response (cooperation or defection) but use contextual cues to infer the decision. The fundamental goal of building trusting relationships with artificial agents is thus nuanced by exploring the boundaries and behavioral consequences of trust and trustworthiness.

In sum, many studies support the idea of intelligent agents as teammates in complex settings and that human-like virtual agents lead to specific affective and behavioral responses of the user. However, it is still unclear when and how trust and cooperation emerge and develop in team-like scenarios that are dynamic, require competence, and extend over many interactions.

## 3 Overview of Approach

We aim to investigate how trust in and cooperation with a human-like agent evolve over time. It is crucial to investigate these dynamics as humans are very sensitive to the social implication of others' behavior for themselves [13]. We assume this is also true when interacting with artificial agents given that people tend to anthropomorphize computers [22]. In line with previous work, we consider computers as social agents that people interact with in a meaningful way and contrast an embodied agent with a non-embodied one. We adopt a dynamic human–agent interaction scenario with a joint goal, allowing us to manipulate key components of cooperation between social agents. This extends prior research using standard cooperative games in that we unravel how humans perceive intelligent agents in strategic problem-solving.

**Interaction Scenario.** We propose an interaction framework in order to analyze key characteristics of social cooperative behavior – human-like cues, trust, competence, trustworthiness – in a systematic fashion. The general setting has two partners solve a puzzle game interactively. Here we present and motivate the general framework and describe the interaction scenario used in the experiment. Inspired by Tetris, the interaction scenario consists of a board where two players work together to place blocks of two shapes, using horizontal movements and rotation. In contrast to Tetris, blocks do not move down gradually and filled lines are not cleared. The latter eases the implementation of an algorithm for the virtual agent to participate as autonomous player. The interaction scenario encompasses a number of actions and elements that can be arranged hierarchically (see Fig. 1). At layer one the agents are both working towards a joint goal, which requires them to coordinate and place the blocks in the puzzle game competently and in response to each others' actions. This layer hence pertains to joint problem-solving based on competence and coordination. The activities on this layer largely determine if the puzzle is solved efficiently or not. Layer two offers the possibility to exchange task-related information. The human players can request this information and need to decide whether the agent's task-related suggestion can be trusted. Trusting a suggestion depends on its quality, the agent's trustworthiness (i.e., competence, perceived intentions), warmth, and other factors we exclude here (e.g., an individual's propensity to trust, own competence). Depending on how much the agent is perceived as social entity, people may see it as tool, assistant, or partner with its own goals. Layer three adds an external goal connected to a specific payoff and hence implies strategic cooperation. In the present study (see below), the payoff is equal for both players. In sum, people may assume the agent is essentially responsive to their actions on the lowest layer, yet how much this commitment comprises support [6] or benevolence [18] is unknown.

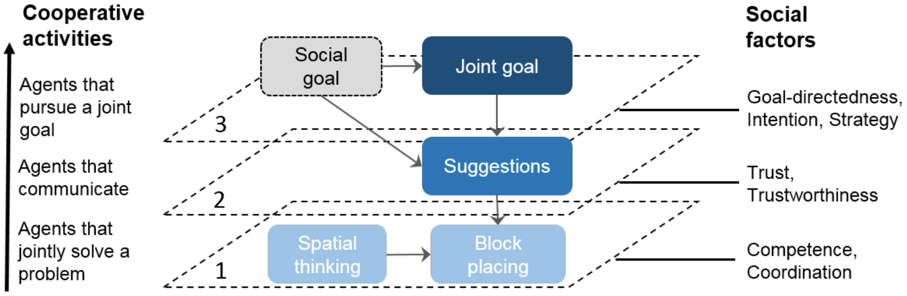
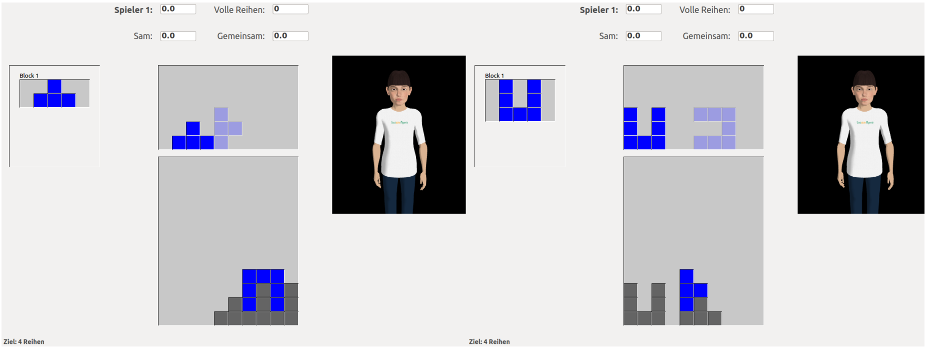


**Fig. 1.** Cooperative activities in the game and corresponding social factors.

## 4 Experiment

We conducted a laboratory experiment where participants tried to solve a puzzle with an embodied vs. non-embodied that offers task-related suggestions, allowing us to directly analyze the effect of human-like cues on trust. To tease apart how much people are willing to trust computer generated advice in situations that are highly interdependent and thus require competence as well as coordination, we also varied the quality of the agent's suggestions. The study had a 2 (agent embodiment: yes vs. no) X 2 (suggestion quality: good vs. bad) between-subjects design with 55 participants ($M_{age} = 25.07$, $SD_{age} = 4.99$) taking part in exchange for 5 EUR.

### 4.1 Method



Spieler 1: 0.0 Volle Reihen: 0

Sam: 0.0 Gemeinsam: 0.0

Spieler 1: 0.0 Volle Reihen: 0

Sam: 0.0 Gemeinsam: 0.0

**Fig. 2.** The game interface in the embodied conditions. Suggestions by the agent are shown at the top in light blue. Left: the agent provides a useful suggestion. Right: the agent makes a bad suggestion.

The game proceeds in turn-based interactions. Players draw one of two available blocks from an urn without replacement. In each round the agent is the first to choose a block, leaving the remaining one to the participant. The joint goal is to complete a specific number of rows such that it is entirely filled with blocks (see Fig. 2). In contrast to Tetris, completed rows are not emptied and there is no time restriction. Completing a row yields 100 points for each player. In each game, the total goal yields a joint payoff such that the score gets doubled for each player. Thus the payoff for both players is always identical. Participants were instructed to work toward the joint goal together with their partner. They were also told that throughout the game, their partner would offer suggestions as to how they could place their block and that they are not obliged to respond in a specific manners. The interaction lasted three games in total with the goal becoming increasingly more difficult (4 rows, 5 rows, 6 rows). The goal was displayed beneath the puzzle field. The progress toward the goal and the payoffs were shown above of it, hence participants saw the distance to the goal and when it was attained. After each game, participants were given a summary sheet showing the payoffs and whether the goal was attained. Before the interaction, participants familiarized themselves with the controls and mechanics without an agent being present. After the interaction, participants filled out the post-questionnaire and rated the agent on task-related social dimensions.

**Manipulations.** The first factor, agent embodiment, determined whether participants played with a virtual agent we called Sam that was introduced as virtual person (*E*: embodied) or with a computer (*NE*: non-embodied). Sam was positioned next to the puzzle field. Aside from eye blinking and breathing behavior, Sam did not show specific nonverbal behaviors. The second factor, suggestion quality, determined how the agent made a suggestion that was drawn from a heuristic we implemented to solve the remaining puzzle field. At each step, the heuristic computes a path to complete the whole field with as few as possible empty fields. Thus the agent suggested either the most (*GS*: good suggestion) or least efficient (*BS*: bad suggestion) solution for the block the human was about to place, such that a bad suggestion would generate empty fields. In each round, the agent would offer three suggestions (in turns 2, 4, 6). In the conditions with embodiment, Sam said *"I have suggestion, do you want to see it?"* or *"I think I know a solution, should I show you?"* Two buttons appeared, labeled "Show me the suggestion" and "I do not want the suggestion", respectively. Thus before the game continued, participants had to decide whether they wanted to see the suggestion or proceed without it. If they decided to see it, a block shape indicated the suggested position and rotation. Importantly, after seeing the suggestion, participants could decide to adopt it or not. In the conditions without embodiment, the buttons appeared at the same locations.

**Dependent Variables.** We segmented the extent to which participants developed trust into three behavioral measures that reflect their response to the agent's offers: offer ignored, suggestion requested but declined, suggestion requested and adopted (for each max. = 9, min. = 0). Second, we computed how often participants attained the goal. Third, participants were asked to rate the agent's trustworthiness and competence, using items proposed by [14] to measure computer credibility (5-point Likert scales). The competence items were 'knowledgeable', 'competent', 'intelligent', 'capable', 'experienced', and 'powerful' (Cronbach's $a = .90$). The trustworthiness items were 'trustworthy', 'good', 'truthful', 'well-intentioned', 'unbiased', and 'honest' ($a = .84$).

**Research Questions.** In this setting the agent offers suggestions like an expert system, yet also plays an active part in the cooperative problem-solving. We hypothesize that the quality of suggestions will impact perceived competence. Second, we explore whether a human-like agent affects people's willingness to request suggestions and their trust in suggestions as indicated by the adoption.

### 4.2 Results

**Goal Attainment.** Figure 3 shows how often the human–agent teams achieved the goal. The maximum per condition was 14 (EGS, EBS) and 13 (NEBS, NEGS), respectively. Again, note that the three games had rising difficulty. We conducted logistic regressions with goal attainment as dependent variable and a) requested suggestion, b) adopted suggestion, c) embodiment and suggestion quality as separate block-wise independent variables, to tease apart whether trust or the conditions influenced goal attainment. No significant effects were found. This indicates that how the agent (and its human partners) actually played was the most important predictor of goal attainment.

**Behavioral Decisions to Trust.** Figure 4 shows participants' responses to the nine agent offerings across all three games. We conducted a 2X2 MANOVA with embodiment and suggestion quality as independent and the three trust variables as dependent variables. The analysis revealed a significant difference for the variables based on the quality of suggestions, Wilk's $\Lambda = .35, F(2, 49) = 46.11, p < .001$ (1

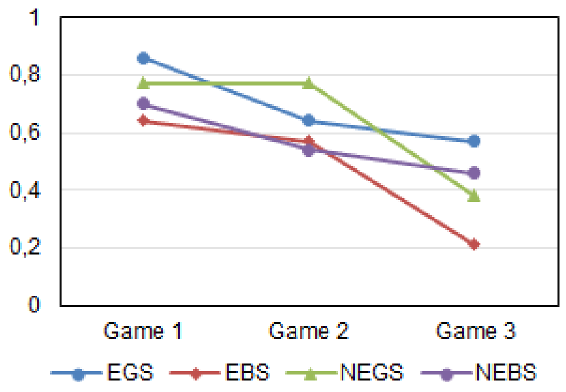
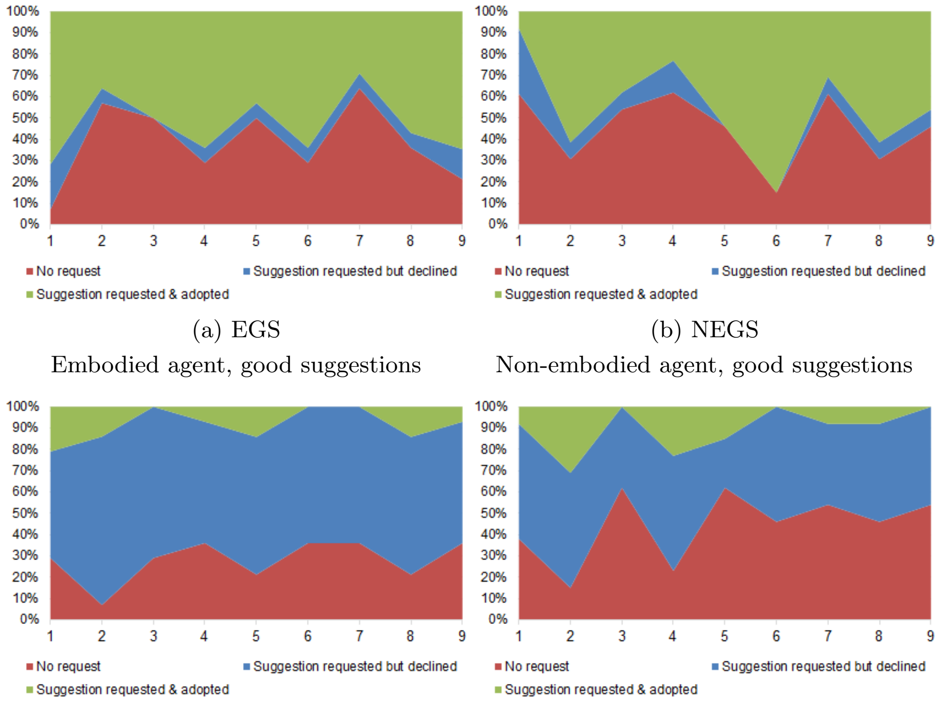


**Fig. 3.** Goal attainment per game with each agent.



(a) EGS
Embodied agent, good suggestions

(b) NEGS
Non-embodied agent, good suggestions

(c) EBS
Embodied agent, bad suggestions

(d) NEBS
Non-embodied agent, bad suggestions

**Fig. 4.** Participant responses to each of the nine agent's offers (suggestion requested or not) and suggestions (adopted or rejected).

case missing). Separate univariate ANOVAs showed that when the suggestion quality was high, participants requested and declined less suggestions, $F(1, 50) = 55.04, p < .001, \eta_p^2 = .52$ ($M = 4.93, SD = 2.70$ vs. $M = .85, SD = 1.11$), and adopted more suggestions, $F(1, 50) = 64.09, p < .001, \eta_p^2 = .56$ ($M = 4.44, SD = 2.02$ vs. $M = .85, SD = 1.01$). Furthermore, there was a tendency among participants' responses showing that when interacting with the embodied agent, they ignored less offers, $F(1, 50) = 3.11, p < .10, \eta_p^2 = .06$ ($M = 2.96, SD = 1.97$ vs. $M = 4.04, SD = 2.46$).

Participants' own decisions determined which elements of the interaction (i.e., good or bad suggestions) they would be exposed to. To help tease apart the effect of embodiment on when a suggestion was requested for the first time, we computed a 2X2 ANOVA with the point at which the first suggestion was requested as dependent variable while ignoring subsequent decisions as, prior to this point, the suggestion quality was unknown. The results revealed that when interacting with the embodied agent, participants requested the first suggestion sooner, $F(1, 50) = 6.20, p < .05, \eta_p^2 = .11$ ($M = 1.21, SD = .50$ vs. $M = 1.81, SD = 1.17$).

**Subjective Ratings.** We conducted a 2X2 MANOVA with embodiment and suggestion quality as independent and perceived competence and trustworthiness as dependent variables. The analysis revealed a significant difference in perceived competence and trustworthiness of the agent based on the quality of its suggestions, Wilk's $\Lambda = .70, F(2, 50) = 10.88, p < .001$. Separate univariate ANOVAs showed that when the suggestion quality was high, the agent was ascribed higher competence, $F(1, 51) = 18.90, p < .001, \eta_p^2 = .27$ ($M = 3.29, SD = .72$ vs. $M = 2.33, SD = .87$), and trustworthiness, $F(1, 51) = 10.76, p < .01, \eta_p^2 = .17$ ($M = 3.35, SD = .82$ vs. $M = 2.60, SD = .85$).

## 5 Discussion

We have presented an experimental design for investigating trust in cooperative human–agent interaction. The results of an experiment conducted within this framework indicate that over time, participants based their decision-making and subjective evaluations of perceived competence and trustworthiness primarily on the quality of suggestions (i.e., competence). However, especially at the beginning of the interaction, the embodied agent clearly facilitated trust in terms of requests for and adoption of suggested actions. It thus seems that while agent embodiment does facilitate initial acceptance and cooperation, this effect does not last. When cooperation needs to extend over a period of time, virtual agents may thus not be displayed constantly but appear when needed. If possible, critical decisions should be addressed at the beginning to leverage the increased level of trust. In our setting, suggestions were not suddenly given but first offered and then provided upon request. When an offer was accepted out of attributed trustworthiness or curiosity, the suggestion could still be rejected. Indeed, the embodied agent evoked the first requested suggestion sooner, indicating a potentially useful effect of human-like cues for cooperation. Further work is needed to investigate the circumstances predicting when this first step is taken by users, and how it could result in trust in terms of advice adoption. The two-step approach may be useful for decreasing regret in decisions and to keep the user in charge. Finally, in a way, our results rectify the social dimension of human–agent cooperation according to which the dynamic process of trusting each other plays an important role. Depending on their own responses, participants took different paths through the interaction. Some trusted the agent early on and presumably assessed the adopted suggestion against their own solutions and competence. Declining an offer right away after suggestions were already requested and/or adopted has thus distinct implications for the trust *quality* and may mean that participants felt the agent's competence would provide no additional value at all. In contrast, requesting but rejecting suggestions has a different meaning as it reflects the need to at least evaluate the agent's competence. This has important implications for how agents should communicate their competence and trustworthiness.

**Acknowledgments.** This research was supported by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster 'its OWL', managed by the Project Management Agency Karlsruhe (PTKA), as well as by the Deutsche Forschungsgemeinschaft (DFG) within the Center of Excellence 277 'Cognitive Interaction Technology' (CITEC).

## References

1. Appel, J., von der Pütten, A., Krämer, N.C., Gratch, J.: Does humanity matter? analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. Advances in Human-Computer Interaction 2012(2), 1–10 (2012)
2. Baylor, A.L.: Promoting motivation with virtual agents and avatars: role of visual presence and appearance. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 364(1535), 3559–3565 (2009)
3. Berry, D.C., Butler, L.T., Rosis, F.d.: Evaluating a realistic agent in an advice-giving task. International Journal of Human-Computer Studies 63(3), 304–327 (2005)
4. Blascovich, J.: A theoretical model of social influence for increasing the utility of collaborative virtual environments. In: Proceedings of the 4th International Conference on Collaborative Virtual Environments, pp. 25–30. ACM, New York (2002)
5. Bradshaw, J.M., Dignum, V., Jonker, C., Sierhuis, M.: Human–agent–robot teamwork. Intelligent Systems, IEEE 27(2), 8–13 (2012)
6. Bratman, M.E.: Shared cooperative activity. The Philosophical Review 101(2), 327 (1992)
7. Breazeal, C., Kidd, C.D., Thomaz, A.L., Hoffman, G., Berlin, M.: Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 708–713. IEEE (2005)
8. Colquitt, J.A., Scott, B.A., LePine, J.A.: Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. Journal of Applied Psychology 92(4), 909–927 (2007)

9. Corritore, C.L., Kracher, B., Wiedenbeck, S.: On-line trust: Concepts, evolving themes, a model. International Journal of Human-Computer Studies 58(6), 737–758 (2003)
10. Dautenhahn, K.: The art of designing socially intelligent agents: Science, fiction, and the human in the loop. Applied Artificial Intelligence 12(7-8), 573–617 (1998)
11. Deutsch, M.: Cooperation and competition. In: Coleman, P.T. (ed.) Conflict, interdependence, and justice, pp. 23–40. Springer, New York (2011)
12. Deutsch, M.: Cooperation and trust: Some theoretical notes. In: Jones, M.R. (ed.) Nebraska Symposium on Motivation, pp. 275–320. University of Nebraska Press, Oxford, England (1962)
13. Fiske, S.T., Cuddy, A.J.C., Glick, P.: Universal dimensions of social cognition: warmth and competence. Trends in Cognitive Sciences 11(2), 77–83 (2007)
14. Fogg, B., Tseng, H.: The elements of computer credibility. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. pp. 80–87. ACM (1999)
15. Hafizoglu, F., Sen, S.: Evaluating trust levels in human-agent teamwork in virtual environments. In *Fourth International Workshop on Human-Agent Interaction Design and Models* (2015), `https://haidm.files.wordpress.com/2015/04/haidm_2015_submission_13.pdf`
16. Kiesler, S., Sproull, L., Waters, K.: A prisoner's dilemma experiment on cooperation with people and human-like computers. Journal of Personality and Social Psychology 70(1), 47 (1996)
17. Krämer, N.C., Rosenthal-von der Pütten, A. M., Hoffmann, L.: Social effects of virtual and robot companions. In: Sundar, S.S. (ed.) The Handbook of the Psychology of Communication Technology, pp. 137–159. John Wiley & Sons (2015)
18. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. The Academy of Management Review 20(3), 709–734 (1995)
19. de Melo, C.M., Carnevale, P., Read, S., Antos, D., Gratch, J.: Bayesian model of the social effects of emotion in decision-making in multiagent systems. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems. pp. 55–62 (2012)
20. Muir, B.M.: Trust between humans and machines, and the design of decision aids. International Journal of Man-Machine Studies 27(5-6), 527–539 (1987)
21. Nass, C., Fogg, B., Moon, Y.: Can computers be teammates? International Journal of Human-Computer Studies 45(6), 669–678 (1996)
22. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. Journal of Social Issues 56(1), 81–103 (2000)
23. Qiu, L., Benbasat, I.: Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. Journal of Management Information Systems 25(4), 145–182 (2009)
24. Traum, D., Marsella, S.C., Gratch, J., Lee, J., Hartholt, A.: Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In: Intelligent Virtual Agents, pp. 117–130. Springer: Berlin, Heidelberg (2008)
25. van Wissen, A., Gal, Y., Kamphorst, B.A., Dignum, M.V.: Human–agent teamwork in dynamic environments. Computers in Human Behavior 28(1), 23–33 (2012)