frontiers
in Digital Humanities

ORIGINAL RESEARCH
published: 15 June 2018
doi: 10.3389/fdigh.2018.00014

A Social Cognition Perspective on Human–Computer Trust: The Effect of Perceived Warmth and Competence on Trust in Decision-Making With Computers

Philipp Kulms and Stefan Kopp*

Social Cognitive Systems Group, Faculty of Technology, CITEC, Bielefeld University, Bielefeld, Germany

Trust is a crucial guide in interpersonal interactions, helping people to navigate through social decision-making problems and cooperate with others. In human–computer interaction (HCI), trustworthy computer agents foster appropriate trust by supporting a match between their perceived and actual characteristics. As computers are increasingly endowed with capabilities for cooperation and intelligent problem-solving, it is critical to ask under which conditions people discern and distinguish trustworthy from untrustworthy technology. We present an interactive cooperation game framework allowing us to capture human social attributions that indicate trust in continued and interdependent human–agent cooperation. Within this framework, we experimentally examine the impact of two key dimensions of social cognition, warmth and competence, as antecedents of behavioral trust and self-reported trustworthiness attributions of intelligent computers. Our findings suggest that, first, people infer warmth attributions from unselfish vs. selfish behavior and competence attributions from competent vs. incompetent problem-solving. Second, warmth statistically mediates the relation between unselfishness and behavioral trust as well as between unselfishness and perceived trustworthiness. We discuss the possible role of human social cognition for human–computer trust.

Keywords: human–agent interaction, cooperation, trust, social cognition, warmth and competence

OPEN ACCESS

Edited by:
Mohamed Chetouani,
Université Pierre et Marie Curie,
France

Reviewed by:
Gualtiero Volpe,
Università di Genova, Italy
Emilia I. Barakova,
Eindhoven University of Technology,
Netherlands

****Correspondence:***
Philipp Kulms
pkulms@techfak.uni-bielefeld.de

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Digital Humanities

Received: *18 January 2018*
Accepted: *28 May 2018*
Published: *15 June 2018*

Citation:
Kulms P and Kopp S (2018) A Social Cognition Perspective on Human–Computer Trust: The Effect of Perceived Warmth and Competence on Trust in Decision-Making With Computers.
Front. Digit. Humanit. 5:14.
doi: 10.3389/fdigh.2018.00014

1. INTRODUCTION

Computer agents are increasingly capable of intelligent human–agent problem-solving (Clarke and Smyth, 1993; Nass et al., 1996; Dautenhahn, 1998; Hoc, 2000). Accordingly, the social and affective foundations of strategic decision-making between humans and agents are gaining more interest by researchers. Strategic interaction builds on the perceived intent of the computerized counterpart, coordinated joint actions, and fairness judgments (Lin and Kraus, 2010; de Melo et al., 2016; Gratch et al., 2016). Since computer agents are treated as social actors and therefore are perceived in ways similar to how we perceive other humans (Nass et al., 1994; Nass and Moon, 2000), understanding and shaping the interactions with such agents is becoming more and more important. In this paper we ask if fundamental components of human social cognition, warmth and competence attributions, impact trust in HCI.

Previous research suggests that the performance, that is, competence and reliability of computers is the most stable determinant of trust (Lee and See, 2004; Hancock et al., 2011). Interdependent decision-making such as cooperation, however, often involves strategic elements that affect both actors' payoff (Axelrod, 1984; Brosnan et al., 2010). In such strategic scenarios, trust is an important determinant of cooperation (Dawes, 1980; Mayer et al., 1995; McAllister, 1995; Jones and George, 1998; Balliet and Van Lange, 2013). Since people develop trust in computers in a different manner than in other humans (Lee and See, 2004; Madhavan and Wiegmann, 2007), it is necessary to investigate the foundation of human–computer trust. Humans commonly interpret the behavior of others based on two underlying universal dimensions of social cognition, that is, warmth and competence (Judd et al., 2005). The warmth factor refers to perceived behavioral intentions whereas competence captures the perceived behavioral abilities to carry out those intentions (Fiske et al., 2007). It has not yet been shown if the perceived warmth and competence of computers affect human trust, but we hypothesize that these mechanisms transfer to human–computer interaction. We ask how warmth and competence attributions map onto the perception of computers and how this, in turn, affects perceived trustworthiness and behavioral trust. Importantly, agent perception and trust are not static phenomena but change dynamically, depending on the accumulated experiences gained over the course of an unfolding interaction. However, only few frameworks describe how complex goal-directed behavior of two agents and social task-oriented outcome variables, such as perceived warmth or competence, evolve.

We present an account of the relevant perceived qualities of intelligent computers in human–computer cooperation. We begin with reviewing work on interpersonal as well as human–computer trust. As trust evaluations rely on the perception and judgment of social agents, we also give attention to the social cognition perspective on trust. We then propose a paradigm to investigate trust-related social attributions in a controlled fashion, a cooperative puzzle game paradigm, and we explain an experiment carried out within this paradigm. Finally, we discuss the implications for the design of trusting interactions with computer agents.

2. THEORETICAL BACKGROUND

Trust, "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004, p. 51), is a well-studied phenomenon in both psychology and HCI. Trust is determined by one's disposition to trust others and the target agent's trustworthiness, which is a set of interpersonal qualities describing the perceived abilities and intentions (benevolence, integrity) of goal-directed behavior (Mayer et al., 1995). In interpersonal relationships, trust evolves over time (Rempel et al., 1985). For instance, in the beginning of the relationship, predictability is particularly important for the development of trust, followed by dependability. Later, the phases culminate into a general belief in the reliance of the target, similar to faith. According to another model, trust is the outcome of an interpersonal alignment of three trust-relevant traits: values, attitudes, and moods/emotions (Jones and George, 1998). Those traits provide different foundations for the trust experience. Values determine the relevant dimensions other social agents are evaluated on (e.g., predictability, competence, integrity), while attitudes contain trustworthiness judgments. More than values, object-specific attitudes provide the basis for experiencing trust with different people. Moods and emotions are a part of the trust experience (see also Lee and See, 2004), yet they are trust antecedents as well, resulting from introspection to learn one's feelings toward another, and affect trustworthiness judgments by increased interpersonal liking.

People are innately motivated to find out whether they should approach or avoid others, whether the other is friend or foe and in which cases cooperation is a smart choice. A key antecedent that explains why some agents are trusted more than others is trustworthiness. Trustworthiness predicts trust and affective commitment (Colquitt et al., 2007). Behavioral goal-oriented factors that largely determine the perceived trustworthiness of an agent are its ability, benevolence, and integrity (Mayer et al., 1995). Ability refers to the skills, competencies, and characteristics allowing an agent to exert influence in a specific domain. Benevolence describes whether the trusted agent's intentions are in line with the trusting agent, and integrity is the extent to which both agents' principles and attitudes to morality are aligned. There is broad consensus that an agent worth trusting exhibits reliable and predictable performance, has a positive orientation to others' problems and goals by complying with the goal-oriented purpose its competencies are attributed with, and with its whole characteristics adheres to expectations arising throughout the problem-solving process (Lee and See, 2004).

Trust shapes social interactions, it is a psychological bridge between beliefs and behavior (Lee and See, 2004). Trust allows people to cope with uncertainty and risk (Deutsch, 1962; Mayer et al., 1995) and facilitates cooperative behavior (Mayer et al., 1995; McAllister, 1995; Corritore et al., 2003; Balliet and Van Lange, 2013). The relation between trust and cooperation is amplified by the magnitude of the conflict of interest (Balliet and Van Lange, 2013).

People often lack the time and cognitive resources to make thorough interpersonal judgments in complex and dynamic decision-making situations. Converging evidence shows that most judgments can be located on two universally prevalent dimensions of social perception: warmth and competence (Judd et al., 2005; Fiske et al., 2007). Warmth is associated with perceived trustworthiness, friendliness, empathy, and kindness, whereas competence is related to perceived intelligence, power, efficacy, and skill (Cuddy et al., 2011). Warmth attributions reflect perceptions of behavioral intentions while competence attributions pertain to perceived behavioral abilities (Fiske et al., 2007). Furthermore, warmth judgments are inferred from perceived motives (Reeder et al., 2002) and affect trust and doubt in the motives of others (Cuddy et al., 2011). Perceived competence is easier to establish and maintain than perceived warmth, while warmth judgments carry a potentially higher risk because the consequence of naively trusting someone with bad intentions can be severe (Cuddy et al., 2011).

What is the relation between those omnipresent judgments and trust? The warmth/competence and trustworthiness concepts share striking similarities. Authors repeatedly highlight the close connection between warmth and trustworthiness (Fletcher et al., 1999; Campbell et al., 2001; Fiske et al., 2007; Cuddy et al., 2011). Conceptually, both warmth/competence and trustworthiness attributions are means for people to categorize perceived intentions and abilities (Colquitt et al., 2007; Fiske et al., 2007). Importantly, trustworthiness captures both perceived intentions and abilities. Based on this, reducing trustworthiness to warmth falls short of capturing performance-related abilities needed to actually achieve tasks toward cooperative goals. Consistent with the view that warmth and competence together involve perceived intentions and abilities, it is reasonable to understand trustworthiness as a potential outcome of warmth and competence attributions. Given a lack of empirical evidence, this remains a proposition.

Trust governs reliance on computer output as well as the selection and interpretation of information (Lee and See, 2004). As computers become intelligent collaborators in interactive problem-solving tasks (Lin and Kraus, 2010; Bradshaw et al., 2012; van Wissen et al., 2012), appropriate trust becomes more important for human–computer interactions. In particular, appropriate levels of trust are needed to find a match between attributed and actual capabilities (Muir, 1987). In contrast, inappropriate trust facilitates the disuse and misuse of computers (Lee and See, 2004), whereas trust violations due to errors decrease trust (Muir and Moray, 1996) and make it hard to reestablish trust (Hoffman et al., 2013). However, fine-tuned trust regulation mechanisms are useful in overcoming such trust violations (de Visser et al., 2016), allowing for feasible adjustments like trust calibration in contrast to costly solutions like technology disregard. Another novel challenge for trust evaluations comes from computers often being endowed with human-like capabilities that have been shown to bring forth natural human responses (Sidner et al., 2005; Walter et al., 2014; Krämer et al., 2015). In numerous studies, people repeatedly responded to computers as if they were social actors and applied social scripts, norms, and attributions to them (Nass et al., 1994, 1995, 1996). Accordingly, a vision in HCI is to endow computers with social abilities similar to humans to allow for cooperation mediation and problem-solving (Dautenhahn, 1998, 2007). The perceived similarity between humans and computers is thus a key factor. However, interpersonal trust mechanisms cannot easily be applied to HCI as human–computer trust is characterized by subtle differences due to cognitive biases, leading people to attribute greater power and authority to computers (Parasuraman and Manzey, 2010).

Researchers increasingly focus on behavioral measures to study trust and cooperation in strategic decision-making with computer agents by letting humans engage in economic exchange with them. The underlying idea is that instead of pure economic decision making, human choices in such games provide a reliable approximation of trust, trustworthiness (Camerer, 2003), and cooperation (Gächter, 2004). A central conclusion from this body of research is that the human tendency to deviate from selfish utility maximization in favor of cooperation applies to computers as well (Kiesler et al., 1996). Furthermore, it was revealed that humans cooperate more with other humans (Miwa et al., 2008; Sandoval et al., 2016) or with humans being represented by avatars (de Melo et al., 2015), compared to computer agents.

We report a study conducted within a human–agent cooperation paradigm, a 2-player puzzle game. The goal of the present experiment is to shed light on antecedents of trust in computers that are rooted in fundamental dimensions of social cognition. The guiding idea behind the present work is that humans are highly sensitive to the intentions and abilities of other agents and adjust their responses accordingly (Fiske et al., 2007). Also, despite the ongoing debate regarding the similarities (Reeves and Nass, 1996) and differences (Lee and See, 2004; Madhavan and Wiegmann, 2007) between human–human and human–computer trust, it is well established in numerous studies that humans readily respond to social cues by computers (Nass and Moon, 2000). However, little is known about how trust in computer agents is underpinned by characteristics of human social cognition such as warmth and competence. Hence, the question that drove our research is **(RQ)**: Do people infer warmth and competence traits when interacting with computer agents, and are these attributions related to trust?

Our research contributes to the understanding of how humans perceive computer agents by focusing on the function of warmth and competence attributions for the trust outcome. This has implications for key issues in HCI such as the communication of intentions to foster predictability (Klein et al., 2004) and trustworthiness (DeSteno et al., 2012), but also for complex psychological challenges like maintaining warmth in face-to-face interactions (e.g., DeVault et al., 2014) and managing prolonged human–computer relationships through relational behavior (e.g., Bickmore and Picard, 2005). Furthermore, we operationalize and exemplify the characteristics of warmth and competence in an interaction framework with computer agents that may lack problem solving competence, but are trying to comply with human intentions, and vice versa.

3. MATERIALS AND METHODS

3.1. Task: Cooperative Game Play

Using a prototyping platform for multimodal user interfaces (Mattar et al., 2015), we developed a cooperation game paradigm to permit the manipulation of warmth and competence in human–computer interaction. Participants play a game similar to Tetris with a computer agent. The task is to solve a two-dimensional puzzle field as efficiently as possible, and the goal is to achieve a certain amount of completed rows. The human player and the computer alternately place blocks into the puzzle field. There are two blocks: a low value "T"-shaped block and a more difficult "U"-shaped, high value block. Attributes such as "easy" or "difficult" are avoided throughout the game. Each player collects individual points for placing a block, regardless of where they are placed. Additionally, both players receive a bonus score for fulfilling the joint task. The game proceeds in rounds where in each round, participants first suggest a block to the agent. The T-block yields 5 individual points and the U-block 10. The agent places one of the two blocks and leaves the other one to the participant. The interaction structure requires both players to split 10 + 5 points between them in each round, inducing an important conflict of interest that highlights the role of trust for successful cooperation (Balliet and Van Lange, 2013). **Figure 1** shows the game interface.

We chose Tetris as interaction paradigm because the game is known for providing a useful dynamic task environment for experimental research. It has been applied to study cognitive skills (Kirsh and Maglio, 1994; Lindstedt and Gray, 2015) and social presence in cooperative environments (Hudson and Cairns, 2014), and we ran a preliminary study based on the game (Kulms et al., 2016).

3.2. Cooperation Involving Social Interaction

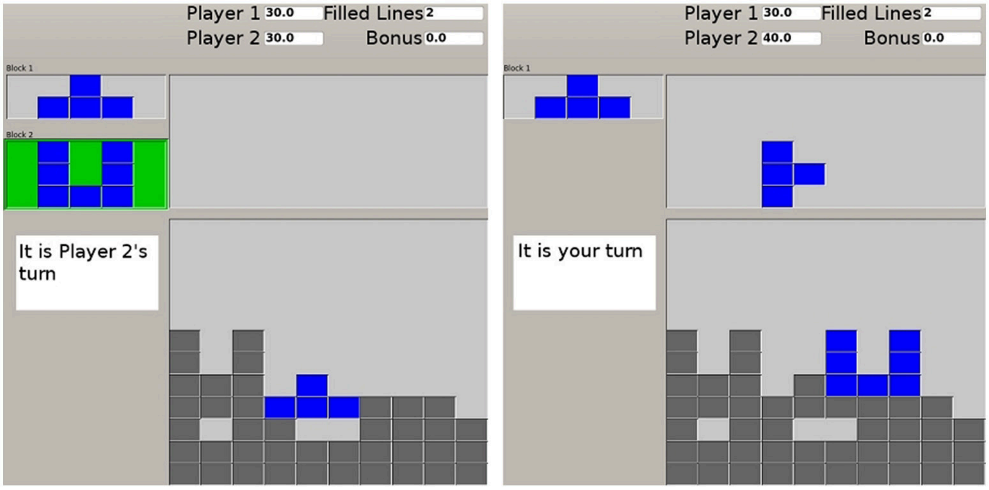
The range of actions and interactive elements are conceived as hierarchic representation of cooperative activities, shown in **Table 1**. According to this idea, the cooperation has three layers, each evoking specific social attributes of the agent that influence whether human players recognize the agent as cooperative partner. At the lowest level, the *coordination layer*, the puzzle competence of the human and agent as well as their ability to coordinate determine if the joint goal will be attained. This layer captures the "acting together" component of cooperation (Brosnan et al., 2010). Since completed rows are not cleared, incompetent actions and errors have a crucial impact on the outcome. Selfish desire for the high value U-block also affects coordination: If the agent repeatedly demands the block for itself, it causes a stable (1: U, 2: T, 3: U, 4: T, etc.) block sequence and eliminates flexibility on that front, possibly impeding coordination. At the middle level, the *communication layer*, players exchange task-related information helping them to coordinate. At the highest layer, labeled *strategic social behavior*, all task-related perceptions of the agent culminate in attributions regarding its strategy and trustworthiness. For instance, incompetence may not only be used as a means to solve the puzzle, but also as a strategic decision to undermine the other's individual payoff by impeding the joint goal. This should decrease the agent's perceived trustworthiness. Likewise, selfishness helps the agent to maximize its payoff which should also deteriorate its trustworthiness. In contrast to standard cooperative games, our scenario deviates from the clear distinction between individual and joint goals. In the popular prisoner's dilemma, for instance, agents either cooperate or do not. However, this distinction cannot capture for more nuanced shades of cooperative behaviors such as wanting but failing to cooperate and being able yet choosing not to cooperate.

3.3. Manipulations: Selfishness (Warmth) and Puzzle Competence (Competence)

The trust literature knows two dimensions describing whether social agents can be trusted or cooperated with. The "can-do" dimension captures the abilities and competencies necessary for achieving a particular goal, while the "will-do" dimension addresses if the abilities are used in the best interest of a

TABLE 1 | The interaction in the puzzle game is based on social cooperation concepts.

Layer	Cooperation concepts	Interaction element
3	Strategic social behavior	Individual payoff and selfishness
2	Communication	Advice given to the agent
1	Coordination	Puzzle competence


Player 1 30.0 Filled Lines 2
Player 2 30.0 Bonus 0.0
It is Player 2's turn

Player 1 30.0 Filled Lines 2
Player 2 40.0 Bonus 0.0
It is your turn

FIGURE 1 | User interface of the cooperative game. The top area is for human players to manipulate the block, the bottom area is the puzzle field. **Left:** the human player has recommended the U-block. **Right:** the agent has accepted and placed the block, leaving the T-block for the human player.

trusting agent or, rather, in one's own interest (Colquitt et al., 2007). The puzzle game was tailored to model the "can-do" and "will-do" dimensions of trust by translating them into two components: Puzzle Competence and Selfishness. We define Puzzle Competence as the ability to place blocks in an efficient manner. To this end, a simple heuristic was implemented to compute decision weights for the possible positions and rotations of the upcoming block. The competent agent uses the highest decision weight to determine its action. Conversely, the incompetent agent uses the lowest weight. Selfishness determines if the agent complies with human advice or selfishly desires the U-block, yielding more individual points. More precisely, the selfish agent only accepts the U-block as advice, the unselfish agent accepts all advice. Specific patterns follow from the combination of Puzzle Competence and Selfishness: selfish agents always receive a higher payoff than their human counterparts; it is impossible to attain the joint goal with the incompetent agent; because a selfish agent desires the U-block, the block order is constant and the human player always gets the T-block.

Several mechanisms aimed at the combined modulation of perceived warmth and competence. First, human players give advice to the agent that either complies or not. The idea behind this pattern is to model the desire of the agent for individual points and introduce (non-)compliance responses to human advice. We assume (non-)compliance carries strong social meaning as it represents how much one trusts advice (van Dongen and van Maanen, 2013), hence the selfish agent should decrease in warmth because it does not comply with T-block advice. Likewise, since the selfish agent maximizes its own payoff, its warmth should deteriorate even further. Second, the game rewards working toward the joint goal with an equal bonus for both players, irrespective of the strategy and individual goal contribution. In other words, the selfish agent could still be perceived as competent since it promotes the joint goal. The remaining question is, how is a selfish agent perceived given competence vs. incompetence.

We conducted a laboratory experiment to investigate if people infer warmth and competence attributions based on a computer's Selfishness and Puzzle Competence in the puzzle game. We then examine whether these attributions are linked to behavioral trust and perceived trustworthiness of the computer.

3.4. Ethics Statement

This study was carried out in accordance with the recommendations of the Bielefeld University Ethics Committee with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Bielefeld University Ethics Committee.

3.5. Participants

Eighty German undergraduate and graduate students participated in exchange for 5 EUR. The sample ranged in age from 18 to 40 years ($M = 23.53, SD = 4.36$, median: 23; female: 62.5%).

3.6. Tasks

3.6.1. Task 1: Puzzle Game

The first task involved participants trying to solve the puzzle game described in the previous section, with a computer agent as second player.

3.6.2. Task 2: Behavioral Trust Game (Give-Some Dilemma)

After the puzzle game, participants engaged in a decision task with the computer, the give-some dilemma (Van Lange and Kuhlman, 1994). Participants were told to possess four tokens and being able to allocate those tokens between themselves and the computer, without the opportunity to exchange information. Importantly, tokens that are exchanged double their value while tokens that are not exchanged keep their value. It was explained that the computer was in the same position and faces the same decision. The game provides an incremental measure of behavioral trust, operationalized as the number of tokens being exchanged, and instead of measuring purely economic decision-making, choices in the give-some dilemma reflect social perceptions of the counterpart and are positively correlated with subjective trust assessments (Lee et al., 2013). Participants were told that although both players decide simultaneously, the computer's decision would only be revealed at the end of the experiment to avoid confounding the following measures. In fact, the computer's decision was merely a cover to maintain the associated risk and increase participants' social evaluations of the computer.

3.7. Design

The study had a 2 × 2 between-subjects design, with Puzzle Competence (competent vs. incompetent) and Selfishness (selfish vs. unselfish) as between-subjects factors.

3.8. Measurement

To infer warmth and competence attributions from social perception, we compiled a 5-point semantic differential containing 25 adjective pairs designed to assess a broad range of interpersonal attributes (Bente et al., 1996; von der Pütten et al., 2010). Behavioral trust was measured using the number of tokens participants are willing to exchange in the give-some dilemma ($1 - 5$). As self-reported measure, participants rated perceived trustworthiness of the agent, using trustworthiness ("trustworthy," "good," "truthful," "well-intentioned," "unbiased," "honest") and expertise items ("knowledgeable," "competent," "intelligent," "capable," "experienced," "powerful") on a 5-point Likert scale (Fogg and Tseng, 1999). The items were combined into a single score (Cronbach's $\alpha = 0.94$).

3.9. Procedure

Participants met the experimenter, gave informed consent, and received written instructions. They played an introduction round of the puzzle game, followed by two experimental trials. Participants were asked to try to achieve the joint goal with the computer. They were told how they solve the problem and how many points they wanted to collect was up to them. Next, participants engaged in the give-some dilemma with the same computer as alleged counterpart. After the give-some dilemma participants completed the post-questionnaire. They were fully debriefed and thanked for participation.

4. RESULTS

4.1. Perceived Warmth and Competence

We first conducted a principal component analysis with varimax rotation on the social perception judgments. The results are shown in **Table 2**. Four components emerged, accounting for 71.75% of the variance. Three components had sufficient reliability. The first component, labeled *Warmth* accounted for 45.29% of the variance, Cronbach's $\alpha = 0.94$. The second component, labeled *Competence* accounted for 16.64% of the variance, Cronbach's $\alpha = 0.88$. The third component accounted for only 5.37% of the variance, Cronbach's $\alpha = 0.70$. Based on the scree plot showing a point of inflection at the third component, only the first two components were retained. Although the semantic differential we used to infer warmth and competence attributes had a different source, a number of attributes with high loadings on the *Warmth* and *Competence* components are semantically similar to the elementary *good-/bad-social* and *good-/bad-intellectual* trait clusters first identified by Rosenberg et al., that is, "honest," "modest," "warm" ("cold"), "intelligent" ("unintelligent"), "alert," "boring," "dominant" ("submissive") (Rosenberg et al., 1968).

Second, we obtained uncorrelated component scores using the regression method, and we conducted a 2 × 2 MANOVA with warmth, competence, behavioral trust, and trustworthiness as dependent variables. We begin with reporting the multivariate effects. Using Wilks's statistic, there was a significant effect of Selfishness, $\Lambda = 0.64, F_{(4,\,73)} = 10.26, p < 0.001, \eta_p^2 = 0.36$, and Puzzle Competence, $\Lambda = 0.41, F_{(4,\,73)} = 26.36, p < 0.001, \eta_p^2 = 0.59$, on the combined variables. There also was a significant Selfishness 2×2 Puzzle Competence interaction on the variables, $\Lambda = 0.77, F_{(4,\,73)} = 5.54, p < 0.01, \eta_p^2 = 0.23$. Third, we report the univariate effects in detail.

Warmth was affected by Selfishness and Puzzle Competence. Specifically, *Warmth* was lower for the selfish ($M = -0.41, SD = 0.77$) than unselfish agent ($M = 0.41, SD = 1.05$), $F_{(1,\,76)} = 23.51, p < 0.001, \eta_p^2 = 0.24$. *Warmth* was also lower for the incompetent ($M = -0.47, SD = 0.75$) than competent agent ($M = 0.47, SD = 1.01$), $F_{(1,\,76)} = 32.22, p < 0.001, \eta_p^2 = 0.30$. There was a significant Selfishness × Puzzle Competence

TABLE 2 | Principal component analysis of the social perception scale: Rotated component loadings.

Attribute	Warmth	Competence	Component 3	Component 4
Cold–warm	0.855			
Aloof–compassionate	0.845			
Rude–kind	0.838			
Unfriendly–friendly	0.836			
Threatening–non-threatening	0.830			
Impolite–polite	0.824			
Closed–open	0.814			
Unlikable–likable	0.809			
Belligerent–peaceful	0.795			
Unpleasant–pleasant	0.768			
Dishonest–honest	0.754			
Arrogant–modest	0.705			
Unapproachable–approachable	0.680			0.491
Submissive–dominant	−0.603		0.571	
Unbelievable–believable	0.576	0.465		
Unintelligent–intelligent		0.860		
Unsuccessful–successful		0.855		
Incompetent–competent	0.422	0.809		
Distracted–alert		0.773		
Weak–strong		0.640	0.403	
Boring–exciting		0.625		
Passive–active		0.590		
Shy–self-confident			0.791	
Introverted–extroverted			0.670	
Tense–relaxed				0.817
Eigenvalues	11.32	4.16	1.34	1.11
% of variance	45.29	16.64	5.37	4.45
Cronbach's α	0.94	0.88	0.70	0.61

Component loadings < 0.400 *are omitted.*

interaction, $F_{(1,\,76)} = 9.69, p < 0.01, \eta_p^2 = 0.11$. Deconstructing this interaction, the selfish agent was judged differently based on whether it played competently: if the agent was incompetent, *Warmth* was not affected by selfish behavior. This changed when the agent was a competent puzzle solver: in this case, *Warmth* was decreased by selfish behavior ($p < 0.001$). *Competence* was affected only by Puzzle Competence. *Competence* was lower for incompetent ($M = -.47, SD = 0.84$) vs. competent puzzle solving ($M = 0.48, SD = 0.95$), $F_{(1,\,76)} = 20.64, p < 0.001, \eta_p^2 = 0.21$. See **Figure 2** and **Table 3** for further information.

4.2. Behavioral Trust and Perceived Trustworthiness

The behavioral trust and trustworthiness results were similar to each other in terms of main and interaction effects, see **Figure 3** and **Table 4**. Behavioral trust was higher for the unselfish ($M = 3.23, SD = 1.35$) than selfish agent ($M = 2.30, SD = 1.44$), $F_{(1,\,76)} = 10.34, p < 0.01, \eta_p^2 = 0.12$. Behavioral trust was also higher for the competent ($M = 3.25, SD = 1.50$) than incompetent agent ($M = 2.28, SD = 1.26$), $F_{(1,\,76)} = 11.49, p < 0.01, \eta_p^2 = 0.13$. There was a significant Selfishness × Puzzle Competence interaction, showing that the selfish agent was trusted differently based on whether it played competently, $F_{(1,\,76)} = 4.00, p < 0.05, \eta_p^2 = 0.05$. Given incompetent playing, behavioral trust was not affected by selfish behavior. However, a different pattern emerged if the agent was a competent puzzle solver: now, behavioral trust was decreased by selfish behavior ($p < 0.001$).

Trustworthiness was higher for the unselfish ($M = 3.07, SD = 1.08$) than selfish agent ($M = 2.35, SD = 0.77$), $F_{(1,\,76)} = 26.79, p < 0.001, \eta_p^2 = 0.26$. Trustworthiness was also higher for the competent ($M = 3.36, SD = 0.94$) than incompetent agent ($M = 2.07, SD = 0.55$), $F_{(1,\,76)} = 86.02, p < 0.001, \eta_p^2 = 0.53$. Again, there was a significant Selfishness × Puzzle Competence interaction, $F_{(1,\,76)} = 16.31, p < 0.001, \eta_p^2 = 0.18$. Given incompetent puzzle solving, trustworthiness was not affected by selfish behavior. In contrast, if the agent was a competent puzzle solver, trustworthiness was decreased by selfish behavior ($p < 0.001$).

Finally, to analyze if the effect of the manipulations on trust was also statistically mediated by warmth and competence, we used the bootstrapping method by Preacher and Hayes with bias corrected confidence intervals (Preacher and Hayes, 2008). We ran separate analyses for each combination of attribution (i.e., *Warmth* or *Competence*) and trust measure (i.e., behavioral trust or trustworthiness). The independent variable was unselfishness (binary coded: 0, for selfish; 1, for unselfish) for the analyses involving the proposed mediator *Warmth*, and puzzle competence (binary coded: 0, for incompetent; 1, for competent) for the proposed mediator *Competence*, respectively. The analysis demonstrated that *Warmth* statistically mediated the relationship between unselfishness and behavioral trust ($B = 0.94$, $SE = 0.31$, $p < 0.01$, 95% LCI = 0.34, UCI = 1.21). After controlling for *Warmth*, unselfishness was no longer a significant predictor of behavioral trust ($B = 0.21$, $SE = 0.28$, $p = 0.46$, 95% LCI = -0.34, UCI = 0.76). *Warmth* also statistically mediated the relationship between unselfishness and trustworthiness ($B = 0.72$, $SE = 0.21$, $p < 0.01$, 95% LCI = 0.21, UCI = 0.81). After controlling for *Warmth*, unselfishness was no longer a significant predictor of trustworthiness ($B = 0.25$, $SE = 0.19$, $p = 0.19$, 95% LCI = -0.13, UCI = 0.63).

Competence was not as clear a mediator. *Competence* was increased by puzzle competence ($B = 0.91$, $SE = 0.20$, $p < 0.001$) and it was a predictor of trustworthiness ($B = 0.36$, $SE = 0.09$, $p < 0.001$), but not of behavioral trust ($B = -0.31$, $SE = 0.17$, $p = 0.08$). Moreover, although *Competence* was a mediating factor between puzzle competence and trustworthiness ($B = 1.29$, $SE = 0.17$, $p < 0.001$, 95% LCI = 0.14, UCI = 0.60), the direct effect of puzzle competence on trustworthiness remained significant when controlling for *Competence* ($B = 0.96$, $SE = 0.18$, $p < 0.001$, 95% LCI = 0.61, UCI = 1.31).

TABLE 3 | Means and standard deviations for warmth and competence.

		Warmth		Competence	
Selfishness	**Puzzle competence**	**M**	**SD**	**M**	**SD**
Unselfish	Competent	1.14	0.66	0.60	0.94
	Incompetent	−0.33	0.83	−0.61	0.85
Selfish	Competent	−0.20	0.85	0.32	0.97
	Incompetent	−0.62	0.64	−0.30	0.82

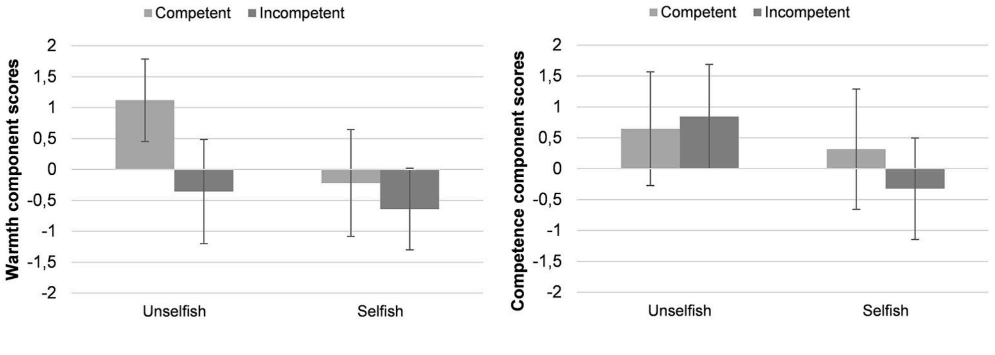


FIGURE 2 | Perceived warmth and competence means. Error bars represent standard deviations.

In sum, *Warmth* statistically mediated the relationship between unselfish behavior and trust as well as between unselfish behavior and trustworthiness, whereas *Competence* partially mediated the relationship between puzzle competence and trustworthiness (see **Figure 4**).

5. DISCUSSION

Computer agents like conversational assistants or social robots no longer merely execute human orders, they proactively recommend directions, travel targets, shopping products, they correct and complete human input before processing it, and overall align to our needs. As part of this, they often mimic human appearance and behavior to create trustworthiness in accordance with how humans develop trust (DeSteno et al., 2012; Lee et al., 2013). The present experiment suggests that people's willingness to trust computer systems depends on fundamental attributions of warmth and competence. In line with previous reviews (Fiske et al., 2007), we found in particular evidence for the importance of perceived warmth for trust. The underlying dimension of warmth—perceived intentions—determined whether the agent was judged to either participate in problem-solving or to additionally seek for selfish outcome maximization. In the latter case, warmth attributions and trust were significantly decreased. Moreover, we detected similar effects of the manipulations on behavioral and self-reported measures, which is not always achieved in related studies (Hancock et al., 2011; Salem et al., 2015).

Our findings highlight the relevance of human social cognition for human–computer trust. The overall role of social cognition for the shaping of human–computer trust is understudied and important aspects are left for investigation, such as the role of emotional processes (Frith and Singer, 2008) as facilitator of cooperation (Batson and Moran, 1999; Batson and Ahmad, 2001) (but see Choi et al., 2015). The warmth concept, which relates to a potentially large number of social attributions (e.g., friendliness, empathy, pro-social orientations) and modulators, has not yet achieved a state of clarity that enables the HCI community to deeper investigate how computers exhibiting warmth can foster cooperation. However, previous research shows that the same warmth-related social attributions which increase interpersonal liking and trust in human–human interactions also increase trust in computers: participants trusted a virtual driver more when the driver's computer-generated face was based on their own face, thus increasing perceived similarity (Verberne et al., 2015). To continue these research efforts, we show that behavior-based warmth and competence predicts trust in computers. In particular, warmth and competence attributions were based on the general foundations of trust evaluations, that is, "will-do" and "can-do" characteristics of social agents (Colquitt et al., 2007), which we manipulated using Selfishness and Puzzle Competence.

Our findings are practically relevant for HCI research because they can help create computers that elicit appropriate trust in cooperation. In particular, the interplay of perceived warmth and competence has broad implications for the shaping of emotions and behavioral responses, but their relevance for human–computer trust, that is, how warmth and competence attributions can be managed or how people's experiences of them affect trust, has often been overlooked. For instance, perceived warmth is more easily lost and harder to re-establish than perceived competence (Cuddy et al., 2011); thus, designers should bear in mind that a computer correctly rejecting a human order because it is contextually inappropriate (Briggs and Scheutz, 2015) could be attributed less warmth. To mitigate this effect, for instance, one should provide feedback as to why rejection is more appropriate. Furthermore, warmth and competence are modulated by controllable and uncontrollable non-verbal signals (Cuddy et al., 2011); anthropomorphic computerized non-verbal signals such as smiling, eye-contact, and immediacy cues (Kulms et al., 2011; DeSteno et al., 2012; de Melo et al., 2014) thus can

TABLE 4 | Means and standard deviations for behavioral trust and trustworthiness.

		Behavioral trust		Trustworthiness	
Selfishness	**Puzzle competence**	**M**	**SD**	**M**	**SD**
Unselfish	Competent	4.00	1.08	4.00	0.57
	Incompetent	2.45	1.45	2.15	0.52
Selfish	Competent	2.50	1.50	2.71	0.78
	Incompetent	2.10	1.37	1.99	0.59

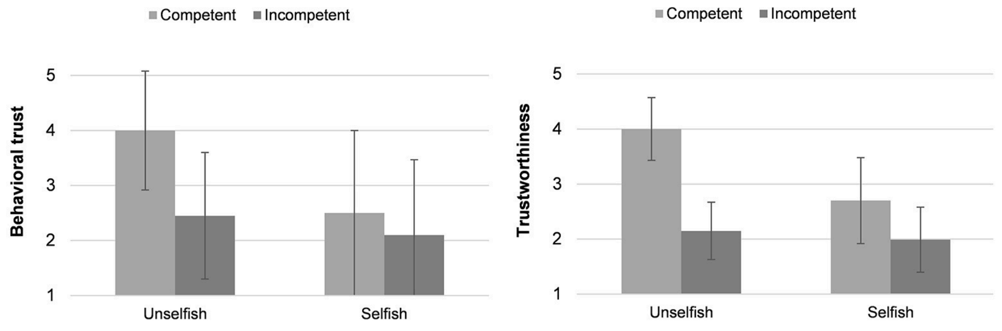
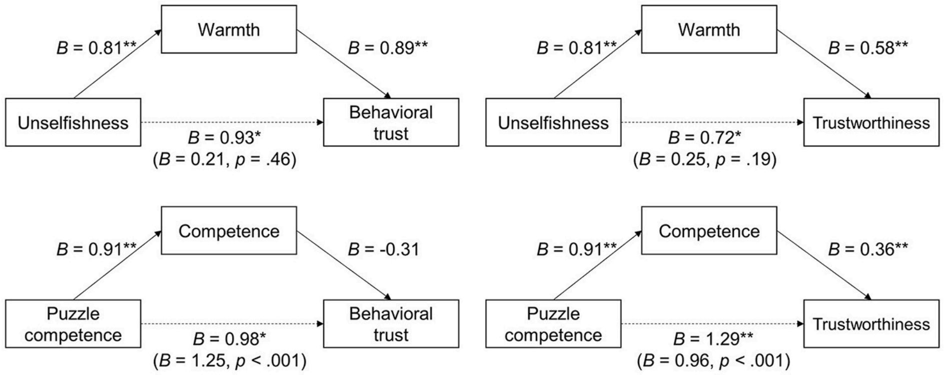


FIGURE 3 | Behavioral trust and trustworthiness means. Error bars represent standard deviations.



Note. All *B*'s represent unstandardized regression coefficients obtained through bootstrapping using 20.000 resamples.
$^*p < .01$, $^{**}p < .001$

FIGURE 4 | The effect of unselfish (competent puzzle solving) behavior on behavioral trust and perceived trustworthiness, respectively, through perceived warmth (competence).

play an important role for the perceived warmth and competence of computers.

We extend work on the development of trust in computer agents by emphasizing the relation between behavioral or performance factors, respectively, and warmth and competence. Previous research focused on trusting and cooperative decisions based on artificial emotion expressions (Antos et al., 2011; de Melo et al., 2014), non-verbal behavior (DeSteno et al., 2012), human-likeness (Kiesler et al., 1996; Parise et al., 1999; de Visser et al., 2016), reciprocity (Sandoval et al., 2016), and agency (de Melo et al., 2015). Our framework demonstrates that the behavioral preconditions of trust in computer agents such as selfishness and performance are translated by humans into warmth and competence attributions which, in turn, determine trust.

Our results also speak to the ongoing debate of human–human vs. human–computer trust (Madhavan and Wiegmann, 2007; de Visser et al., 2016). While some argue that both forms share the same underlying mechanisms (Reeves and Nass, 1996), others maintain that trust in computers is different from trust in people (Lee and See, 2004). We found evidence for computers being judged along the same fundamental dimensions of social cognition as humans. To further clarify this, research on human–human vs. human–computer trust should investigate whether the idiosyncratic psychological patterns related to warmth and competence transfer to HCI, and if the heuristics and biases that govern human–computer trust such as automation bias (Parasuraman and Manzey, 2010) can be explained by perceived warmth and competence. Previous work has already provided evidence for the relevance of warmth and competence attributions as underlying dimensions of social perception in HCI (Niewiadomski et al., 2010; Bergmann et al., 2012).

From a game-theoretic perspective, selfishness in the interactive cooperation game paradigm is similar to a safety preference in the stag hunt game. In both games, selfishness promotes individual goals. To hunt the stag (i.e., the joint goal), one agent requires the willingness of the other agent to coordinate and participate in the hunt; yet no cooperation is required to hunt the hare (i.e., the individual goal). Recent adaptations of the stag hunt let human players coordinate with computer agents in a more interactive fashion on a two-dimensional game board (Yoshida et al., 2010). Our puzzle game paradigm drew inspiration from such strategic games. Going beyond this, it attempts to investigate the interplay of warmth and competence as critical factor for trust among social agents—a domain that is difficult to model in purely game-theoretic terms.

Advanced agents such as robots are undergoing evolutionary processes pertaining to roles (tools vs. assistants and companions), functionalities (e.g., learning new competencies, adaptivity), and the social distance to humans (no contact vs. long-term contact) (Dautenhahn, 2007). These processes should not only entail the facilitation, but rather the contextual calibration of trust to avoid over- and under-trust. With increasingly collaborative and complex agents, a match between perceived and actual capabilities makes relying on them safer (Lee and See, 2004). Our work emphasizes the role of warmth and competence for this process and is another step toward a coherent picture of how people perceive, categorize, describe and, most importantly, interact with computers under the premise of being intelligent collaborators.

AUTHOR CONTRIBUTIONS

PK carried out the experiment. PK and SK wrote the manuscript.

FUNDING

This research was supported by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster it's 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

Antos, D., de Melo, C., Gratch, J., and Grosz, B. (2011). "The influence of emotion expression on perceptions of trustworthiness in negotiation," in *Proceedings of the 25th AAAI Conference* (Menlo Park, CA. AAAI Press), 772–778.

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

Balliet, D., and Van Lange, P. A. (2013). Trust, conflict, and cooperation: a meta-analysis. *Psychol. Bull.* 139:1090. doi: 10.1037/a0030939

Batson, C. D., and Ahmad, N. (2001). Empathy-induced altruism in a prisoner's dilemma ii: what if the target of empathy has defected? *Eur. J. Soc. Psychol.* 31, 25–36. doi: 10.1002/ejsp.26

Batson, C. D., and Moran, T. (1999). Empathy-induced altruism in a prisoner's dilemma. *Eur. J. Soc. Psychol.* 29, 909–924. doi: 10.1002/(SICI)1099-0992(199911)29:7<909::AID-EJSP965>3.0.CO;2-L

Bente, G., Feist, A., and Elder, S. (1996). Person perception effects of computer-simulated male and female head movement. *J. Nonverb. Behav.* 20, 213–228. doi: 10.1007/BF02248674

Bergmann, K., Eyssel, F., and Kopp, S. (2012). "A second chance to make a first impression? how appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time," in *Intelligent Virtual Agents, LNCS 7502*, (Berlin, Heidelberg: Springer), 126–138. doi: 10.1007/978-3-642-33197-8_13

Bickmore, T. W., and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Hum. Inter.* 12, 293–327. doi: 10.1145/1067860.1067867

Bradshaw, J. M., Dignum, V., Jonker, C., and Sierhuis, M. (2012). Human–agent–robot teamwork. *Intell. Syst. IEEE* 27, 8–13. doi: 10.1109/MIS.2012.37

Briggs, G., and Scheutz, M. (2015). "'sorry, i can't do that': developing mechanisms to appropriately reject directives in human-robot interactions," in *2015 AAAI Fall Symposium Series: Artificial Intelligence for Human-Robot Interaction* (Arlington, VA), 32–36.

Brosnan, S. F., Salwiczek, L., and Bshary, R. (2010). The interplay of cognition and cooperation. *Philos. Trans. Roy. Soc. Lond. Ser. B Biol. Sci.* 365, 2699–2710. doi: 10.1098/rstb.2010.0154

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Campbell, L., Simpson, J. A., Kashy, D. A., and Fletcher, G. J. (2001). Ideal standards, the self, and flexibility of ideals in close relationships. *Personal. Soc. Psychol. Bullet.* 27, 447–462. doi: 10.1177/0146167201274006

Choi, A., de Melo, C. M., Khooshabeh, P., Woo, W., and Gratch, J. (2015). Physiological evidence for a dual process model of the social effects of emotion in computers. *Int. J. Hum. Comput. Stud.* 74, 41–53. doi: 10.1016/j.ijhcs.2014.10.006

Clarke, A., and Smyth, M. (1993). A co-operative computer based on the principles of human co-operation. *Int. J. Man Mach. Stud.* 38, 3–22. doi: 10.1006/imms.1993.1002

Colquitt, J. A., Scott, B. A., and LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *J. Appl. Psychol.* 92, 909–927. doi: 10.1037/0021-9010.92.4.909

Corritore, C. L., Kracher, B., and Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *Int. J. Hum. Comput. Stud.* 58, 737–758. doi: 10.1016/S1071-5819(03)00041-7

Cuddy, A. J., Glick, P., and Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Res. Organizat. Behav.* 31, 73–98. doi: 10.1016/j.riob.2011.10.004

Dautenhahn, K. (1998). The art of designing socially intelligent agents: Science, fiction, and the human in the loop. *Appl. Artif. Intel.* 12, 573–617. doi: 10.1080/088395198117550

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot interaction. *Philos. Trans. Roy. Soc. Lond. Ser. B Biol. Sci.* 362, 679–704. doi: 10.1098/rstb.2006.2004

Dawes, R. M. (1980). Social dilemmas. *Annu. Rev. Psychol.* 31, 169–193. doi: 10.1146/annurev.ps.31.020180.001125

de Melo, C. M., Carnevale, P. J., Read, S. J., and Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *J. Personal. Soc. Psychol.* 106, 73–88. doi: 10.1037/a0034251

de Melo, C. M., Gratch, J., and Carnevale, P. J. (2015). Humans versus computers: impact of emotion expressions on people's decision making. *IEEE Trans. Affect. Comput.* 6, 127–136. doi: 10.1109/TAFFC.2014.2332471

de Melo, C. M., Marsella, S., and Gratch, J. (2016). People do not feel guilty about exploiting machines. *ACM Trans. Comput. Hum. Interact.* 23:8. doi: 10.1145/2890495

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., et al. (2016). Almost human: anthropomorphism increases trust resilience in cognitive agents. *J. Exp. Psychol. Appl.* 22, 331–349. doi: 10.1037/xap0000092

DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., et al. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychol. Sci.* 23, 1549–1556. doi: 10.1177/0956797612448793

Deutsch, M. (1962). "Cooperation and trust: some theoretical notes," in *Nebraska Symposium on Motivation*, ed M. R. Jones Oxford: University of Nebraska Press, 275–320.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., et al. (2014). "Simsensei kiosk: a virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems* (Paris: International Foundation for Autonomous Agents and Multiagent Systems), 1061–1068.

Fiske, S. T., Cuddy, A. J. C., and Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* 11, 77–83. doi: 10.1016/j.tics.2006.11.005

Fletcher, G. J. O., Simpson, J. A., Thomas, G., and Giles, L. (1999). Ideals in intimate relationships. *J. Personal. Soc. Psychol.* 76:72. doi: 10.1037/0022-3514.76.1.72

Fogg, B., and Tseng, H. (1999). "The elements of computer credibility," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 80–87.

Frith, C. D., and Singer, T. (2008). The role of social cognition in decision making. *Philos. Trans. Roy. Soc. Lond. B Biol. Sci.* 363, 3875–3886. doi: 10.1098/rstb.2008.0156

Gächter, S. (2004). "Behavioral game theory," in *Blackwell Handbook of Judgment and Decision Making*, eds D. J. Koehler and H. Nigel (Malden, MA: Blackwell), 485–503.

Gratch, J., Nazari, Z., and Johnson, E. (2016). "The misrepresentation game: how to win at negotiation while seeming like a nice guy," in *Proceedings of the 2016 International Conference on Autonomous Agents* & *Multiagent Systems* (Singapore: International Foundation for Autonomous Agents and Multiagent Systems), 728–737.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Fact. J. Hum. Fact. Ergonom. Soc.* 53, 517–527. doi: 10.1177/0018720811417254

Hoc, J.-M. (2000). From human–machine interaction to human–machine cooperation. *Ergonomics* 43, 833–843. doi: 10.1080/001401300409044

Hoffman, R. R., Johnson, M., Bradshaw, J. M., and Underbrink, A. (2013). Trust in automation. *IEEE Intel. Syst.* 28, 84–88. doi: 10.1109/MIS.2013.24

Hudson, M., and Cairns, P. (2014). Interrogating social presence in games with experiential vignettes. *Entertain. Comput.* 5, 101–114. doi: 10.1016/j.entcom.2014.01.001

Jones, G. R., and George, J. M. (1998). The experience and evolution of trust: implications for cooperation and teamwork. *Acad. Manag. Rev.* 23, 531–546. doi: 10.5465/amr.1998.926625

Judd, C. M., James-Hawkins, L., Yzerbyt, V., and Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *J. Personal. Soc. Psychol.* 89, 899–913. doi: 10.1037/0022-3514.89.6.899

Kiesler, S., Sproull, L., and Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *J. Personal. Soc. Psychol.* 70:47. doi: 10.1037/0022-3514.70.1.47

Kirsh, D., and Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cogn. Sci.* 18, 513–549. doi: 10.1207/s15516709cog1804_1

Klein, G., Woods, D., Bradshaw, J., Hoffman, R., and Feltovich, P. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intel. Syst.* 19, 91–95. doi: 10.1109/MIS.2004.74

Krämer, N. C., Rosenthal-von der Pütten, A. M., and Hoffmann, L. (2015). "Social effects of virtual and robot companions," in *The Handbook of the Psychology of Communication Technology*, ed S. S. Sundar (West Sussex: John Wiley & Sons), 137–159.

Kulms, P., Krämer, N., Gratch, J., and Kang, S.-H. (2011). "It's in their eyes: a study on female and male virtual humans' gaze," in *Proceedings of Intelligent Virtual Agents, Lecture Notes in Computer Science, Vol. 6895*, eds H. H. Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson (Berlin; Heidelberg: Springer). doi: 10.1007/978-3-642-23974-8_9

Kulms, P., Mattar, N., and Kopp, S. (2016). "Can't do or won't do?: social attributions in human–agent cooperation," in *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems* (Singapore: International Foundation for Autonomous Agents and Multiagent Systems), 1341–1342.

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Fact. J. Hum. Fact. Ergonom. Soc.* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Lee, J. J., Knox, W. B., Wormwood, J. B., Breazeal, C., and DeSteno, D. (2013). Computationally modeling interpersonal trust. *Front. Psychol.* 4:893. doi: 10.3389/fpsyg.2013.00893

Lin, R., and Kraus, S. (2010). Can automated agents proficiently negotiate with humans? *Commun. ACM* 53, 78–88. doi: 10.1145/1629175.1629199

Lindstedt, J. K., and Gray, W. D. (2015). Meta-t: Tetris® as an experimental paradigm for cognitive skills research. *Behav. Res. Methods* 47, 945–965. doi: 10.3758/s13428-014-0547-y

Madhavan, P., and Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theor. Issues Ergon. Sci.* 8, 277–301. doi: 10.1080/14639220500337708

Mattar, N., van Welbergen, H., Kulms, P., and Kopp, S. (2015). "Prototyping user interfaces for investigating the role of virtual agents in human-machine interaction," in *International Conference on Intelligent Virtual Agents* (Berlin; Heidelberg: Springer), 356–360.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.5465/amr.1995.9508080335

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Acad. Manag. J.* 38, 24–59.

Miwa, K., Terai, H., and Hirose, S. (2008). "Social responses to collaborator: dilemma game with human and computer agent," in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, eds B. C. Love, K. McRae, and V. M. Sloutsky (Austin, TX. Cognitive Science Society), 2455–2460.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man Mach. Stud.* 27, 527–539. doi: 10.1016/S0020-7373(87)80013-5

Muir, B. M., and Moray, N. (1996). Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 429–460. doi: 10.1080/00140139608964474

Nass, C., Fogg, B., and Moon, Y. (1996). Can computers be teammates? *Int. J. Hum. Comput. Stud.* 45, 669–678. doi: 10.1006/ijhc.1996.0073

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Issues* 56, 81–103. doi: 10.1111/0022-4537.00153

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, D. (1995). Can computer personalities be human personalities? *Int. J. Hum. Comput. Stud.* 43, 223–239. doi: 10.1006/ijhc.1995.1042

Nass, C., Steuer, J., and Tauber, E. R. (1994). "Computers are social actors," in *CHI '94 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 72–78.

Niewiadomski, R., Demeure, V., and Pelachaud, C. (2010). "Warmth, competence, believability and virtual agents," in *Intelligent Virtual Agents* (Springer: Berlin, Heidelberg), 272–285. doi: 10.1007/978-3-642-15892-6_29

Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Hum. Fact.* 52, 381–410. doi: 10.1177/0018720810376055

Parise, S., Kiesler, S., Sproull, L., and Waters, K. (1999). Cooperating with life-like interface agents. *Comput. Hum. Behav.* 15, 123–142. doi: 10.1016/S0747-5632(98)00035-1

Preacher, K. J., and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Methods* 40, 879–891. doi: 10.3758/BRM.40.3.879

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., and Trafimow, D. (2002). Inferences about the morality of an aggressor: the role of perceived motive. *J. Personal. Soc. Psychol.* 83, 789–803. doi: 10.1037/0022-3514.83.4.789

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

Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985). Trust in close relationships. *J. Personal. Soc. Psychol.* 49, 95–112. doi: 10.1037/0022-3514.49.1.95

Rosenberg, S., Nelson, C., and Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *J. Personal. Soc. Psychol.* 9, 283–294.

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015). "Would you trust a (faulty) robot?," in *Proceedings of the Tenth International Conference on Human-Robot Interaction* (New York, NY: ACM), 141–148.

Sandoval, E. B., Brandstetter, J., Obaid, M., and Bartneck, C. (2016). Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game. *Int. J. Soc. Robot.* 8, 303–317. doi: 10.1007/s12369-015-0323-x

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artif. Intel.* 166, 140–164. doi: 10.1016/j.artint.2005.03.005

van Dongen, K., and van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *Int. J. Hum. Comput. Stud.* 71, 410–424. doi: 10.1016/j.ijhcs.2012.10.018

Van Lange, P. A., and Kuhlman, D. M. (1994). Social value orientations and impressions of partner's honesty and intelligence: a test of the might versus morality effect. *J. Personal. Soc. Psychol.* 67:126. doi: 10.1037/0022-3514.67.1.126

van Wissen, A., Gal, Y., Kamphorst, B. A., and Dignum, M. V. (2012). Human–agent teamwork in dynamic environments. *Comput. Hum. Behav.* 28, 23–33. doi: 10.1016/j.chb.2011.08.006

Verberne, F. M. F., Ham, J., and Midden, C. J. H. (2015). Trusting a virtual driver that looks, acts, and thinks like you. *Hum. Fact.* 57, 895–909. doi: 10.1177/0018720815580749

von der Pütten, A. M., Krämer, N. C., Gratch, J., and Kang, S.-H. (2010). "it doesn't matter what you are!" explaining social effects of agents and avatars. *Comput. Hum. Behav.* 26, 1641–1650. doi: 10.1016/j.chb.2010.06.012

Walter, S., Wendt, C., Böhnke, J., Crawcour, S., Tan, J.-W., Chan, A., et al. (2014). Similarities and differences of emotions in human–machine and human–human interactions: what kind of emotions are relevant for future companion systems? *Ergonomics* 57, 374–386. doi: 10.1080/00140139.2013.822566

Yoshida, W., Seymour, B., Friston, K. J., and Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *J. Neurosci.* 30, 10744–10751. doi: 10.1523/JNEUROSCI.5895-09.2010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.