It's (Not) Your Fault! Blame and Trust Repair in Human-Agent Cooperation

Victoria Buchholz * **Philipp Kulms** * **Stefan Kopp** *

* *Social Cognitive Systems Group, Center of Excellence 'Cognitive Interaction Technology' (CITEC), Faculty of Technology, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany (e-mail: {vbuchholz, pkulms, skopp}@techfak.uni-bielefeld.de)*

Abstract: In cooperative settings the success of the team is interlinked with the performance of the individual members. Thus, the possibility to address problems and mistakes of team members needs to be given. A common means in human-human interaction is the attribution of blame. Yet, it is not clear how blame attributions affect cooperation between humans and intelligent virtual agents and the overall perception of the agent. In order to take a first step in answering these questions, a study on cooperative human-agent interaction was conducted. The study was designed to investigate the effects of two different blaming strategies used by the agent in response to an alleged goal achievement failure, that is, self-blame (agent blames itself) followed by an apology versus other-blame (agent blames the user). The results indicate that the combination of blame and trust repair enables a successful continuation of the cooperation without loss of trust and likeability.

Keywords: Human-agent interaction; trust; cooperation; blame

1. INTRODUCTION

Over the years the nature of human-computer interaction changed from using the machines as mere tools to viewing them as potential team mates in cooperative settings (Bradshaw et al., 2012; Nass et al., 1996). People already interact with human-like virtual agents similarly to how they would with other humans (see Krämer et al., 2015, for an overview). Therefore, not only humans, but also intelligent computer agents should be endowed with means to address problems and mistakes in cooperative settings. In human-human interaction a common and effective way to achieve this is blame. Blame attributions provide the opportunity to regulate behaviour, identify problems, and help fixing them (Malle et al., 2014; Groom et al., 2010). Identifying under which conditions humans accept blame attributions from a computer agent and trust the agent's judgement is crucial. Research in human-robot interaction already addressed the topic of blame and identified disadvantages, e. g. the decrease of trust (Groom et al., 2010; Kaniarasu and Steinfeld, 2014). Therefore, one key aspect for attributing blame in human-computer cooperation is that trust in an agent is repaired successfully after a blame judgement. However, it is not clear if blame affects human-agent interaction in the same way as it affects human-robot interaction and how trust can be repaired successfully. In this paper we take a first step in investigating the effects of blame and trust repair in human-agent interaction in a specific cooperative setting. In particular, by conducting a human-agent cooperation study, a first attempt is made to examine under which circumstances the advantages of blame can be applied appropriately such that the cooperation can continue without loss of trust and acceptance.

We begin with reviewing the theoretical background and related work on blame, trust and trust repair in section 2. In section 3 we give an overview of our approach by describing the used framework and a pre-study. In section 4 we then present results from the human-agent cooperation study, which are subsequently discussed in section 5.

2. BACKGROUND AND RELATED WORK

In a cooperative game, where "a team member's outcome is tied to the outcome of the entire team" (Nass et al., 1996, 671), people get a feeling of affiliation regardless whether the members of the team are humans or computers. Recent evidence also shows that humans and machines are able to work together in various settings and that cooperation is an important ability of intelligent systems (Bradshaw et al., 2012). The concept of blame is not new to human-machine interaction. Moon (2003), for example, found that people often assign responsibility according to the self-serving bias when interacting with computers. This bias denotes the tendency to claim credit for a success and blame the partner for a failure. So far, however, the use of blame in human-agent cooperation has not been investigated.

2.1 Blame

Malle et al. (2014) define blame as a form of public criticism which can only be directed at a person, relies on social cognition and always demands warrant. This strict definition makes it possible to distinguish between blame and wrongness judgements, anger and event evaluations. Wrongness judgements target an intentional behaviour,

event evaluations an event and anger can be directed at anything. All three do not require warrant. Moreover, in comparison to anger, blame is not an emotion people can feel and directly express. It is a form of responsibility assignment which requires information processing. The purpose of blaming is the regulation of social behaviour. It is a powerful but also dangerous means of communication (Malle et al., 2014). Hence, the blamer needs to carefully consider how he or she approaches the target of blame. This can be difficult whenever the accompanying emotions are strong and overwhelming. Yelling or personally attacking the interlocutor would not be helpful, whereas "people welcome thoughtful, clear, and constructive criticism" (Voiklis et al., 2014, 1701). Blaming can be seen as an invitation to communicate. When the invitation is accepted, both sides can explain themselves, the relationship can be repaired, and the blamer has the possibility to persuade the norm violator to change his or her behaviour (Malle et al., 2014). Accordingly, if blame is attributed in a thoughtful way, it is a useful means to communicate problems and to regulate behaviour.

The emotion accompanying blame does not only depend on whether we blame ourselves or another person. It also depends on the controllability or intentionality of the behaviour in question. When we blame another person we most likely feel anger or pity: Anger when we perceive the behaviour as intentional or controllable and pity when we judge the behaviour as unintentional or uncontrollable (Sander and Scherer, 2009). Similarly, we most probably feel guilty when we intentionally commit or omit to do something, whereas we feel ashamed when the negative outcome was not controllable. Hence, the expression of blame varies, as blame can be accompanied by different emotions, e. g. anger, pity, guilt, and shame. In this experiment the focus lies on intentional or controllable blame and therefore the expressions of guilt and anger will be looked at in more detail.

Whereas anger is a distinct facial expression, guilt is not (Keltner and Buswell, 1996). Guilt, as well as the emotion of regret, appear when one is responsible for a negative outcome that affects others and not only oneself (Zeelenberg and Breugelmans, 2008). However, Keltner and Buswell (1996) propose that people communicate guilt in a different way than just with their face. Likewise, anger is not exclusively associated with blame and therefore insufficient to communicate blame alone. Thus, for this experiment the facial expressions of anger for other-blame and regret for self-blame are used in connection with the verbal expression of blame.

At this point, we are not aware of any research on the effects of blame attribution in human-agent interaction, but studies investigating this topic in human-robot interaction already exist. Groom et al. (2010) and Kaniarasu and Steinfeld (2014), for example, examined the effects of three different types of blame attributions by robots: self-, other- and team-blame. Their results demonstrated that humans prefer self-blame over other-blame in robots and even like the self-blaming robot better than the team-blaming robot. Other-blame only showed negative effects on the perception of the robot's abilities and the feeling of team affiliation. Additionally, participants in the other-blame condition perceived the robot as less competent. Regarding their results on trust, there was no significant difference between the types of attributions. Accordingly, blame lowers trust, either because of the users' annoyance of the robot that blames them or because the users perceive a constantly apologising robot as not trustworthy (Kaniarasu and Steinfeld, 2014).

2.2 Trust Violations and Trust Repair

Lee and See (2004) define trust as the expectation of an agent (the trustor) that another agent (the trustee) will help to achieve the trustor's goals in an uncertain and vulnerable situation. Trust violations occur when the trustee does not fulfil the expectation of the trustor. Typically, trust violations lead to three emotions: anger, disappointment, and "regret over having trusted in the first place" (Martinez and Zeelenberg, 2015, 119). In an experiment Martinez and Zeelenberg (2015) found contrasting effects of these emotions: While anger and regret decrease trust and trustworthiness, disappointment increases them. But trust is not only affected by the feelings of the trustor, it is also influenced by the feelings of the trustee. When the trustee feels responsible for the harm done to the trustor, he or she blames him- or herself. Hence, feelings of regret and guilt accompanying self-blame (Kim et al., 2006) decrease trust. Furthermore, they could lead to a confession or an apology in order to make up for the misbehaviour (Keltner and Buswell, 1996) and repair the relationship. Kim et al. (2006) found evidence of positive effects of different types of apologies after trust violations that are competence- or integrity-based. Successful trust repair after integrity-based trust violations is achieved by apologising with an external attribution rather than an internal. Consequently, the trust-repairing agent needs to apologise for the misbehaviour, assure that it will not happen again, and partly take responsibility, yet also attribute blame to an external influence, for example to another agent. In comparison, internal attributions are more effective after competence-based trust violations. Accordingly, trust is repaired more successfully if the violator takes full responsibility and apologises for what happened. In addition to the implications of apologies for human-human interaction, a study by de Visser et al. (2016) demonstrates that anthropomorphism enhances the effect of a trust-repairing apology of a computer agent in cooperation.

To summarise, blame emerges when a person is found responsible for a negative event or a norm-violating action and there is reason to believe that he or she acted intentionally or had the obligation and capacity to prevent the outcome (Malle et al., 2014). Furthermore, the attribution of blame and accompanying emotions can have a positive or negative effect on trust and trustworthiness (Kaniarasu and Steinfeld, 2014; Martinez and Zeelenberg, 2015). In case of negative consequences, possibilities to repair trust lie in the form of the apology (Kim et al., 2006). The state of the art demonstrates that human-computer interaction and especially interaction between humans and human-like virtual agents can be social and similar to human-human interaction (Lee and Nass, 2010). Studies focusing on human-robot interaction reveal that self-blame is preferred over other-blame, but all blaming strategies decrease trust (Groom et al., 2010; Kaniarasu and Steinfeld, 2014). Yet, it

is not clear how attributions of blame and trust-repairing apologies affect human-agent cooperation.

3. OVERVIEW OF APPROACH

3.1 Interaction Framework

The cooperative puzzle game was implemented with the MultiPro prototyping framework (Mattar et al., 2015). In the game two players — a human player and the virtual agent — try to attain a joint goal interactively. The puzzle game is inspired by Tetris, but includes some modifications in order to facilitate controllability over the interaction and enable the computer to take part in the game as an autonomous player. There are two different types of blocks, a U-shaped (U-block) and a T-shaped block (T-block). Furthermore, blocks do not fall down gradually, but can be moved horizontally and rotated in 90 degree units at the top of the game board, before they are placed in the puzzle field. The players take turns at placing blocks, thereby trying to fill complete horizontal lines. Filled lines are counted, but not cleared as in Tetris. The joint goal is presented prior to the beginning of the game. In the present study the joint goal is to reach a top 10 list by filling more complete horizontal lines as the previous players. Thereby, the outcome of one player depends on the outcome of the entire team, which may result in a feeling of team affiliation (Nass et al., 1996). The block-placing algorithm used by the agent tests all possible options and searches for the best next move based on different criteria, e. g. whether a line can be filled with this move. The algorithm was slightly adjusted for this experiment, in the form that only the U-block was considered for the next move of the agent instead of both blocks. Due to the agent's frequent use of the U-block, filling lines became somewhat difficult and a tendency to fail was already evident during the game. Moreover, the top 10 list was designed such that all participants failed to reach the joint goal. In this setting, the agent is given more control as it always makes the first move and places the U-block in the puzzle field. Thereafter, the human partner can only take the remaining block and is therefore not given the chance to choose a block. This gives the human partner warrant to blame the agent for a failure in the game. Once the last possible block has been moved into the puzzle field, the round ends and the players get to know whether the goal was attained. The virtual agent is positioned on the right of the puzzle field. In the experiment the agent was referred to as Sam. Figure 1 shows the interface and the agent.

3.2 Pre-Study

A pre-study was conducted ($N = 29$) to examine the assignment of responsibility by the human after (not) attaining the joint goal. The agent only showed eye blinking and breathing behaviour and did not interact or communicate with the participants in any way. The results showed a strong tendency to affiliate with the team: 76% of all participants assigned responsibility to both players instead of themselves or the agent (see Figure 2).

In order to investigate if this changes when the agent actively blames itself or the partner after not attaining the joint goal, a self-blaming agent versus an agent blaming the human partner were used in the main study, which is described in detail in section 4. As ratings of the agent's perceived competence did not differ significantly between the two conditions in the pre-study, it was assumed that participants did not believe the agent brought about the failure due to a lack of competence. Hence, we expected a possible trust violation in the main experiment not to be competence-, but rather integrity-based. With this in mind, successful trust repair after an integrity-based trust violation is achieved by apologising with an external attribution (Kim et al., 2006).

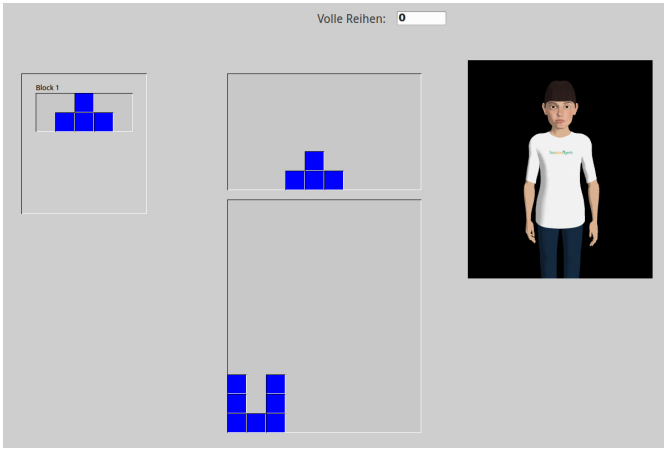

Fig. 1. The game interface with the virtual agent Sam. The next block can be modified in the block field (top) and placed into the puzzle field (bottom).

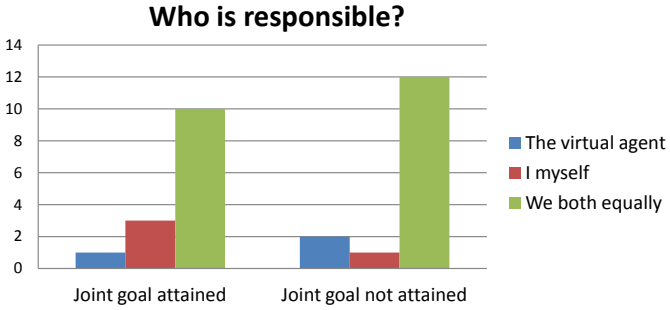

Fig. 2. Responsibility attributions (pre-study).

4. EXPERIMENT

In a laboratory experiment participants played a cooperative puzzle game with a virtual agent. The study employed a between-subject design manipulating the blaming strategies used by the agent after not attaining the joint goal: self-blame followed by a trust-repairing apology (SB condition) and other-blame (OB condition). The effects of the two strategies on the likeability, perceived competence of, and trust in the agent were assessed. It was expected that the two blaming strategies affect the overall perception of the agent differently.

4.1 Participants

Thirty-six participants (20 males and 16 females) between the age of 20 and 34 took part in the study ($M_{age} = 23.86$,

SD_{age} = 2.81). The majority of them were students from Bielefeld University of different study programmes. The participants' experience with the original Tetris game ranged from "non at all" to "much" (M_{tetris} = 3.33, SD_{tetris} = 1.12, evaluated on a five-point Likert scale). The study lasted approximately 30 minutes and the participants took part in exchange for chocolate. Subjects were randomly assigned to the two test groups.

4.2 Procedure

Participants were informed that they took part in an experiment on natural human behaviour in an interactive puzzle game. They were also informed that the second player was controlled by a computer. Participants were asked to sign a consent form and to fill out a pre-questionnaire that covered personal information like age, gender and an estimation of their experience with the game Tetris. Next, participants received the puzzle game instructions containing a description of the course of the game and the keyboard controls needed to move, rotate and place the blocks. Then, participants played an introductory round without the virtual agent to familiarise themselves with the controls and the game interface. Before the interaction with the agent began, the joint goal: "Manage to reach the top 10 list together by filling more lines!" and the top 10 list containing the names of players and the numbers of filled lines they reached in the game were presented on screen. Afterwards, one round of the game was played together with the agent. Once the game ended, the participants saw their result. In order to evaluate the participants' trust in the agent after the game, they played one round of the Give-Some Dilemma described in section 4.3. Subsequently, participants evaluated the interaction with the virtual agent in a post-questionnaire. Upon completing the questionnaire, participants were fully debriefed and received chocolate as a reward for their participation.

4.3 Material

Cooperative Puzzle Game. The game as described in section 3.1 was used and the behaviour of the virtual agent was manipulated. In each of the two conditions the virtual agent Sam reacted at the beginning of the round, after moves five and seven — either after its own move (SB) or after the team partner's move (OB) — as well as at the end of the round after the message "You did not reach the top 10 list!" was shown. Sam's introductory sentence and behaviour did not differ in the two conditions. Here, Sam smiled a little while saying "Hi, I am Sam. I will place the first block now!". The introductory sentence was included in order to inform participants that the agent would communicate with them in the further course of the game and to make the interaction seem more natural. To increase the effect of the blame judgements and to achieve more naturalness in the interaction, the agent reacted twice during the round and once at the end. In the SB condition Sam shook its head and said "What am I doing?" after its fifth move and looked sad and a bit regretful after its seventh move (see Figure 3) while saying "That way we will not achieve the goal!".

Fig. 3. The agent's facial expressions in the SB condition. Left: looking sad and somewhat regretful. Right: showing regret.

At the end of the round the agent tried to repair trust with an external apology. Hence, it apologised, blamed itself partly, while also attributing blame to a virus and stating that it will not happen again. In the game this was implemented by letting Sam show regret (see Figure 3) while saying: "Oh no. Now I am somehow responsible for our failure! I was not able to concentrate today because I have caught a virus. I am sorry! That will not happen again!".

Similarly, in the OB condition the agent shook its head and said: "What are you doing?" after the subject's fifth move and looked a bit angry after the subject's seventh move (see Figure 4) while saying: "That way we will not achieve the goal!". At the end of the round Sam showed anger (see Figure 4) and said: "Now it is your fault that we lost! If you had placed the blocks differently, we would have played better together!".

Fig. 4. The agent's facial expressions in the OB condition. Left: looking somewhat angry. Right: showing anger.

Trust Measure. Trust in the agent was evaluated using the Give-Some Dilemma (GSD), a two-player social dilemma between the participant and the computer agent. In this dilemma both players receive four coins. Their task is to decide how many coins they want to exchange to their partner. Importantly, the value of coins being exchanged is doubled, whereas the value of coins that are kept stays the same. As they make this decision simultaneously, there is no possibility to agree on a strategy. The individual payoff is maximized by receiving *and* keeping all coins, whereas the collective payoff is maximized by both players exchanging all coins, hence the dilemma. The dilemma provides an incremental measure of behavioural trust, operationalised as the number of tokens being exchanged. Instead of measuring pure economic decision-making, choices in the dilemma reflect social perceptions of the counterpart and were shown to correlate positively with subjective trust assessments (Lee et al., 2013). To illustrate how the

game works an example round and a payoff matrix were printed on a sheet for the participants. A one-shot version of the dilemma was played, so that participants needed to make their decisions based solely on their perception of Sam after the puzzle game. The dilemma was included in the game interface to create the impression of an ongoing interaction with Sam, as the agent did not actually take part in the game.

Post-Questionnaire. Participants were asked to evaluate their interaction with Sam in the puzzle game on multiple scales. First, participants were asked to rate their emotional reaction, using the following items (Rilling et al., 2008): jealousy, guilt, betrayal, indignation, anger, contempt, disappointment, envy, happiness, sadness, relief, annoyance, camaraderie, trust, and shame (5-point Likert scale). The same items were used to examine what participants thought Sam felt after the game to assess how well the intentions were conveyed by the facial expressions. A manipulation check assessed participants' understanding of the responsibility assignment: "Who did Sam blame for the result?" (answer options: "Sam", "me" or "us both equally"). To investigate responsibility attributions, participants answered the question "According to your opinion, who is responsible for the outcome?" (answer options: "Sam", "me" or "us both equally"). The question "According to your opinion, how much did Sam try to achieve the joint goal?" assessed the agent's perceived cooperativeness (5-point Likert). The overall experience in the game was evaluated by asking "How much would you like to play again with Sam?" (5-point Likert). Moreover, we measured Sam's perceived competence and trustworthiness. We adopted items proposed by Fogg and Tseng (1999) to measure computer credibility (5-point Likert). The competence items were "knowledgeable", "competent", "intelligent", "capable", "experienced" and "powerful" (Cronbach's $\alpha = .93$); The trustworthiness items were "good", "well-intentioned", "honest", "truthful", "trustworthy" and "unbiased" (Cronbach's $\alpha = .86$). As trustworthiness is a factor of likeability (cf. Fiske et al., 2007), the overall experience and the perceived trustworthiness indicated how much participants liked the virtual agent. Finally, an open question provided the possibility to describe Sam in one's own words.

4.4 Hypotheses

Since there is no comparable research in human-agent interaction and previous studies suggest that humans may behave in the same way with robots as with virtual agents (Lee and Nass, 2010), it is assumed to observe the same results as Kaniarasu and Steinfeld (2014) and Groom et al. (2010) in their experiments on blame attribution in human-robot interaction. Hence, we assume that participants prefer the self-blaming agent over the agent blaming its partner, resulting in higher likeability and competence ratings of the self-blaming agent. As other-blame had a negative effect on team affiliation, we expect that the assignment of responsibility differs significantly between the conditions. Since the self-blaming agent tries to repair trust at the end of the game according to the findings of Kim et al. (2006), we hypothesise that trust in the other-blaming agent is lower than in the self-blaming agent due to the trust-repairing apology.

4.5 Results

Only perceived competence and perceived trustworthiness were normally distributed. Thus the independent t-test was performed for these variables and the Mann-Whitney U test for the not normally distributed variables. The independent variable in both tests was the condition the participants were assigned to. To test for an association between the test groups and the assignment of responsibility the chi-square test was applied.

Manipulation Checks. The manipulation was successful since all participants correctly answered who was blamed by Sam. Furthermore, the ratings of the agent's emotional reaction after the game imply that the facial expressions of regret and anger were perceived as intended. In the SB condition participants rated Sam's guilt higher ($M_{SB} = 4.17$, $SD_{SB} = .79$, $M_{OB} = 1.44$, $SD_{OB} = .98$, $U = 15.00$, $z = -4.90$, $p < .001$, $r = -.82$), whereas participants in the OB group rated anger higher ($M_{SB} = 2.17$, $SD_{SB} = 1.10$, $M_{OB} = 4.00$, $SD_{OB} = .97$, $U = 39.00$, $z = -4.02$, $p < .001$, $r = -.67$).

Likeability and Competence. The perceived trustworthiness differed significantly between the two conditions: Sam was perceived more trustworthy in the SB than in the OB condition ($M_{SB} = 3.81$, $SD_{SB} = .63$, $M_{OB} = 2.55$, $SD_{OB} = .66$, $t(34) = 5.83$, $p < .001$, $r = .70$). There was a significant relationship between this variable and the overall experience, which represented whether the participants would like to play again with Sam ($\tau = .34$, $p < .01$). That suggests that as the ratings of Sam's trustworthiness increases, how much participants would like to play again with Sam increases as well. Moreover, ratings of the overall experience differed significantly between the two conditions. Subjects in the SB group would much rather play with Sam again than subjects in the OB group ($M_{SB} = 4.00$, $SD_{SB} = .97$, $M_{OB} = 2.33$, $SD_{OB} = 1.41$, $U = 60.00$, $z = -3.30$, $p < .01$, $r = -.55$). Another significant correlation was detected between the overall experience and the participants' feeling of camaraderie ($\tau = .38$, $p < .01$). Ratings of the agent's competence did not differ significantly between the test groups ($M_{SB} = 3.12$, $SD_{SB} = .81$, $M_{OB} = 2.90$, $SD_{OB} = .89$, $t(34) = .78$, ns, $r = .13$).

Cooperativeness, Responsibility and Trust. Ratings of the agent's cooperativeness did not differ significantly between the test groups ($M_{SB} = 3.72$, $SD_{SB} = 1.02$, $M_{OB} = 3.39$, $SD_{OB} = 1.14$, $U = 135.00$, $z = -.89$, ns, $r = -.15$). Furthermore, no significant association was found between the conditions and the attribution of responsibility ($\chi^2(2) = 3.05$, ns). Fifty-three percent of all subjects assigned responsibility to the team (see Figure 5).

Ninety-four percent of the subjects explained their reasons for assigning responsibility. In the SB condition 80% of the subjects attributing responsibility to both players felt a team affiliation and thus blamed themselves and Sam for the outcome of the game. Only 67% named the same reason in the OB condition for blaming the team. Additionally, 22% in the OB condition said that, while their mistakes were decisive for the failure, Sam's comments during the game were demotivating and thereby had an effect on the outcome as well. Of the other

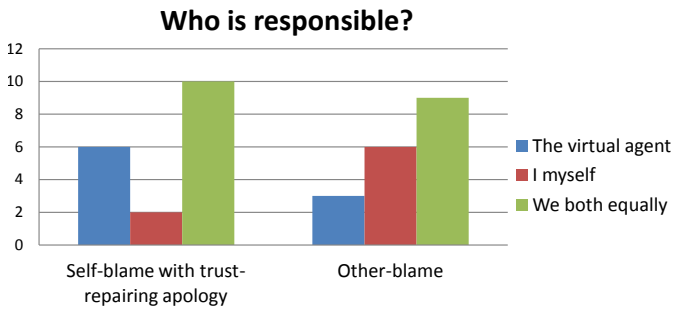

Fig. 5. Responsibility attributions (main study).

47%, however, who did not assign responsibility to both players, 71% agreed with Sam on the attribution of blame regardless of the test group. However, a tendency could be seen in the number of coins given to Sam in the GSD. Subjects in the SB group tended to give more coins to Sam than subjects in the other group ($M_{SB} = 3.47$, $SD_{SB} = 1.60$, $M_{OB} = 2.50$, $SD_{OB} = 1.30$, $U = 86.50$, $z = -1.80$, $p < .10$, $r = -.31$).

Emotional Reaction. The evaluation of the subjects' emotional reaction after the game reveals significant differences between the two test groups. While subjects in the OB condition felt more shame, annoyance, indignation and relief, subjects in the SB condition felt more camaraderie. Furthermore, a significant tendency was observed implying that subjects in the SB condition were happier and felt less guilt and anger. The feeling of trust did not differ significantly between the conditions. Table 1 contains the statistical values for the emotional reaction.

Open Question Comments. Similarities were found between the answers of the participants to the open questions at the end of the post-questionnaire. The participants were asked to describe Sam in their own words. Fifteen percent observed that the agent said nothing positive during the interaction and 12% wished to have more communication in the form of discussing a strategy, getting positive feedback as well as negative or getting suggestions for improvement from Sam.

5. DISCUSSION

We have presented an experiment including a virtual agent that actively attributed blame, either to itself or to a human team partner, in the context of cooperation. As it was assumed that any blaming strategy reduces trust, the self-blaming agent used an external apology in order to repair trust. The results were in line with findings in human-robot interaction concerning the higher likeability of the self-blaming agent (Groom et al., 2010; Kaniarasu and Steinfeld, 2014), as the self-blaming agent was perceived more trustworthy and the participants would much rather play with this agent again than with the other agent. However, in contrast to the results of Groom et al. (2010), ratings of the agent's competence did not differ significantly, which is a positive result for the use of other-blame.

We observed a tendency to trust the self-blaming agent more than the agent blaming the human partner. Participants' feeling of trust, however, did not differ significantly between the two test groups. Nevertheless, the correlation between perceived trustworthiness, whether participants would like to play again with Sam (overall experience), and participants' feeling of camaraderie, might imply that whenever the agent is perceived more trustworthy, feelings of camaraderie and the overall experience are positively influenced. This is a strong indicator for the success of trust repair. Nevertheless, further investigation of self-blame with and without a trust-repairing apology is needed to verify this assumption. So far, it is not clear, if self-blame lowers trust and the repair was successful or if the blaming strategy did not have the same negative effects in this study compared to the experiments with robots by Groom et al. (2010) and Kaniarasu and Steinfeld (2014). A control condition with a self-blaming agent that does not try to repair trust could have confirmed the assumptions.

Moreover, participants' emotional reactions in the SB condition were more positive than in the OB condition. Participants who were blamed by the agent felt more shame, annoyance, indignation and also relief, but less camaraderie. Since relief is a positive emotion, it is not clear how it fits in with the negative emotions those participants felt more strongly. An explanation could be that these participants were relieved that the game and thus the agent's negative comments were over. Moreover, a tendency emerged indicating that participants in the OB condition were angrier, felt more guilt, and were less happy in comparison. Therefore, it seems that the results are in line with the findings of the blame experiments conducted with robots by Groom et al. (2010) and Kaniarasu and Steinfeld (2014).

Participants remarked that the agent should not only give negative but also positive feedback and also make suggestions for improvement in order to keep up the participants' motivation. This might also attenuate any negative effects of blame judgements and especially of those directed at others. Moreover, users would have liked to discuss a strategy with the agent prior to a round. This may improve the cooperative character of the puzzle game.

In contrast to the assumption that other-blame has a negative effect on the team affiliation, a similar effect in the assignment of responsibility was observed in the main study as in the pre-study. Approximately half of the participants assigned responsibility to the team and 80% explained this with the feeling of a team affiliation. This suggests that either the agent's attribution of blame did not have any effect, or participants did not think that one player alone had the capacity to prevent the outcome and therefore attributed less or even no blame. Participants' answers to the open questions suggest that they did believe the agent, but either felt a strong team affiliation with the self-blaming agent or not only blamed themselves, but also the demotivating statements of the agent blaming them. However, as the precise course of the game was different for every participant, the manipulation of the agent's game play might not have been optimal to control how serious participants perceived the agent's mistakes. Minor mistakes might not warrant blame. A suggestion for further experiments thus is that the agent makes more serious mistakes at specified times during the game. As team affiliation seemed to have a strong influence on the assignment of responsibility in both experiments, it would be interesting to compare the SB and OB conditions to a

Table 1. Mann-Whitney U test results of the emotional reactions.

Variable	M_{SB}	SD_{SB}	M_{OB}	SD_{OB}	**U**	**z**	**Sig.**	**r**
shame	1.72	1.07	2.56	1.29	100.00	-2.08	$p < .05$	-.35
annoyance	2.11	1.32	3.11	1.32	95.50	-2.20	$p < .05$	-.37
indignation	1.89	.90	3.06	1.21	76.00	-2.82	$p < .01$	-.47
relief	1.94	.73	2.67	.91	89.50	-2.42	$p < .05$	-.40
camaraderie	2.83	1.15	1.72	1.02	75.50	-2.85	$p < .01$	-.48
happiness	2.67	.97	2.06	.94	104.00	-1.93	$p < .10$	-.32
guilt	1.94	1.11	2.78	1.44	106.00	-1.88	$p < .10$	-.31
anger	1.61	.85	2.22	1.11	110.50	-1.74	$p < .10$	-.29
trust	2.61	1.15	2.06	.94	116.00	-1.51	*ns*	-.25

third in which the agent attributes blame to the team as it was done in the experiments by Groom et al. (2010) and Kaniarasu and Steinfeld (2014).

As participants showed a strong tendency to affiliate with the team in both conditions and ratings of the agent's cooperativeness did not differ significantly, both strategies of blame enable an ongoing interaction. However, blame directed at the user elicited negative responses as well. Hence, other-blame should only be used by the agent if the human team partner is obviously or likely more culpable. In contrast, self-blame with a trust-repairing apology is an effective means for addressing problems, as the cooperation can continue and trust in and likeability of the agent are high. Trust repair after blame attributions may also provide a solution to avoid negative responses to an agent blaming the user.

In conclusion, the results indicate that if blame is attributed correctly, it can be used successfully to address problems and to help fixing them. Implementing trust repair after attributions of blame is likely to be the key for that. Overall, this work has taken a first step in exploring blame attributions and trust repair in human-agent cooperation, but further experiments are needed to fully understand the complex concepts of blame and trust repair.

ACKNOWLEDGEMENTS

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

Bradshaw, J.M., Dignum, V., Jonker, C., and Sierhuis, M. (2012). Human-agent-robot teamwork. *IEEE Intelligent Systems*, 27(2), 8–13.

de Visser, E.J., Monfort, S.S., McKendrick, R., Smith, M.A., McKnight, P.E., Krueger, F., and Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331–349.

Fiske, S.T., Cuddy, A.J., and Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2), 77–83.

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

Groom, V., Chen, J., Johnson, T., Kara, F.A., and Nass, C. (2010). Critic, compatriot, or chump?: Responses to robot blame attribution. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot interaction*, 211–218. ACM, New York.

Kaniarasu, P. and Steinfeld, A.M. (2014). Effects of blame on trust in human robot interaction. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 850–855. IEEE Conference Publications.

Keltner, D. and Buswell, B.N. (1996). Evidence for the distinctness of embarrassment, shame, and guilt: A study of recalled antecedents and facial expressions of emotion. *Cognition and Emotion*, 10(2), 155–171.

Kim, P.H., Dirks, K.T., Cooper, C.D., and Ferrin, D.L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, 99(1), 49–65.

Krämer, N., Rosenthal-von der Pütten, A., and Hoffmann, L. (2015). Social effects of virtual and robot companions. In S. Sundar (ed.), *The Handbook of the Psychology of Communication Technology*, 137–159. John Wiley & Sons, Chichester, UK.

Lee, J.D. and See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1), 50–80.

Lee, J.E.R. and Nass, C.I. (2010). Trust in computers: The Computers-Are-Social-Actors (CASA) paradigm and trustworthiness perception in human-computer communication. In D. Latusek and A. Gerbasi (eds.), *Trust and Technology in a Ubiquitous Modern Environment: Theoretical and Methodological Perspectives*, 1–15. Information Science Reference, Hershey.

Lee, J.J., Knox, W.B., Wormwood, J.B., Breazeal, C., and DeSteno, D. (2013). Computationally modeling interpersonal trust. *Frontiers in Psychology*, 4, 893.

Malle, B.F., Guglielmo, S., and Monroe, A.E. (2014). A theory of blame. *Psychological Inquiry*, 25(2), 147–186.

Martinez, L.F. and Zeelenberg, M. (2015). Trust me (or not): Regret and disappointment in experimental economic games. *Decision*, 2(2), 118–126.

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 W. Brinkman, J. Broekens, and D. Heylen (eds.), *Intelligent Virtual Agents*, 356–360. Springer International

Publishing.

Moon, Y. (2003). Don't blame the computer: When self-disclosure moderates the self-serving bias. *Journal of Consumer Psychology*, 13(1-2), 125–137.

Nass, C., Fogg, B.J., and Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, 45(6), 669–678.

Rilling, J.K., Goldsmith, D.R., Glenn, A.L., Jairam, M.R., Elfenbein, H.A., Dagenais, J.E., Murdock, C.D., and Pagnoni, G. (2008). The neural correlates of the affective response to unreciprocated cooperation. *Neuropsychologia*, 46(5), 1256–1266.

Sander, D. and Scherer, K. (2009). *Oxford Companion to Emotion and the Affective Sciences*. Oxford University Press.

Voiklis, J., Cusimano, C.J., and Malle, B.F. (2014). A social-conceptual map of moral criticism. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 1700–1705. Cognitive Science Society.

Zeelenberg, M. and Breugelmans, S.M. (2008). The role of interpersonal harm in distinguishing regret from guilt. *Emotion*, 8(5), 589–596.