Is an apology enough? How to resolve trust breakdowns in episodic online interactions

Asimina Vasalou  Astrid Hopfensitz  Jeremy Pitt

Electrical and Electronic Engineering  Swiss Center for Affective Sciences  Electrical and Electronic Engineering
Imperial College  University of Geneva  Imperial College
SW7 2BT, London, United Kingdom  1205, Geneva, Switzerland  SW7 2BT, London, United Kingdom
a.vasalou@imperial.ac.uk  astrid.hopfensitz@cisa.unige.ch  j.pitt@imperial.ac.uk

ABSTRACT
This paper addresses what kind of system allows the victim of a trust breakdown to fairly assess an unintentional offender who is also a benevolent member. Two systems were compared: a system that displayed the offender’s unblemished reputation score as obtained in previous interactions with other members, and a system that also had a communication channel which displayed the offender’s expressed apology and regret over the offence. The findings of this study suggest that the system which also endorses apology, as well as records reputation, allows the victim to recover his/her trust in the unintentional offender. However, trust is repaired only when the offender validates the apology with a reparative action.

Categories and Subject Descriptors
G3.2. Computer Communication Networks: Distributed systems
F.2.2. Nonnumerical Algorithms and Data Structures: Artificial intelligence

General Terms
Design, Economics, Experimentation.

Keywords
Apology, Forgiveness, Trust, Reputation, Repair.

1. INTRODUCTION
Researchers have considered anti-normative behaviour in computer-mediated communication (CMC) from two different angles. One stream of research has been concerned with how to engender trust, thus preventing trust breakdowns. For instance, it has been shown that when placed in a social dilemma, members are more likely to cooperate with others over rich media such as video or audio as opposed to being pseudonymous [1]. In online auctioning, buyers can choose which sellers to trust with a transaction based on sellers’ historical behaviour as captured by reputation systems [6]. Media theorists on the other hand, have researched trust breakdowns usually by regarding the offender’s behaviour as a frequent and intentional act. For example, “trolls” are seemingly genuine members who often post taunting messages in newsgroup forums, in truth meant to spark disputes by challenging the group’s beliefs [2]. In this paper we identify a topic that has not been accounted for in these two existing approaches: a trust breakdown may occur as a result of an unintentional offence of an otherwise benevolent member of the community. In everyday life, such offences are inevitable but not unforgivable. Apologies for one can evoke empathy towards the offender and in turn facilitate the process of forgiveness [4]. But in anonymous, episodic interactions which are part of many online settings the resolution of such offences can become increasingly complex. This is partly because the narrow timeframe of each interaction combined with the impoverished communication channel constrict the cues of trustworthiness one can acquire on another member (e.g. identity, integrity, willingness to comply to institutions, benevolence; [8]). At the onset, this establishes interactions that are perceived as more risky, thus building barriers that may stand in the way of resolution if trust breaks down. Even more, reputation systems which have been widely employed to sustain trust in episodic interactions are not necessarily able to repair trust. Reputation systems convey members’ ability and performance, hence all the rational reasons for initiating a transaction in an online auction or for accepting a member’s advice in an online forum. Yet, when trust breaks down, it is not known if a static reputation score can motivate the victim of an offence to forgive, while at the same time indicating the offender’s good intention and willingness to repair. In a recent effort to address this problem, eBay launched a feature called the “mutual feedback withdrawal” by which users can contest the reputation score they received. Only if both the victim and the offender agree to engage in this process, the resolution of the issue is taken offline. Later, the victim may retract and improve the offender’s online reputation score. At heart, eBay offers an online offender an outlet through which to apologize, elaborate on his/her intentions and repair, subsequently allowing the victim of the offence to restore the trust by removing the original low reputation rating.

This paper contrasts a reputation system to a reputation system that also includes an apology channel, parallel to the eBay approach. The aim of this research is to reveal which of the two systems may motivate the victim of a trust breakdown to forgive and once again to trust an unintentional and infrequent offender during a episodic interaction.

2. A DEFINITION OF TRUST FOR CMC
The work presented in this paper was informed by a definition of trust proposed by [7][8]. This definition describes trust in CMC as an asymmetric and asynchronous relationship between two parties, the trustor and the trustee. Asymmetry results as the trustor only assumes risk in anticipation of an asynchronous fulfillment from the trustee. This relationship can be broken down into two sequential moves. In the first move, the trustor assesses the trustee’s perceived trustworthiness. On the basis of this judgment, the trustor may or may not decide to take some risk over an interaction that can have an uncertain outcome. From his part, the trustee, who is the second mover, may either fulfill or violate the first mover’s trust. This high level definition applies to numerous online interactions each of which may
present different types of risk. For example, members of online emotional support forums post private information expecting empathy in return. In online auctioning, buyers pay before receiving the goods.

3. STUDY
3.1 Overview
The study conducted captured the asynchronous and asymmetric definition of trust with the trust game [7] which has been used as a research paradigm in previous CMC studies (e.g. [9]). The trust game assigns one player to be first mover (trustor) while a second player is the second mover (trustee). The payoffs of the game we used were taken from [3]. The first mover starts with 150 points, 50 points more than the second mover. The first mover may choose to transfer his extra 50 points to the second mover or to withdraw from the risk. If the first mover transfers the points, they are multiplied by 6 before they are given to the second mover. The second mover has the option to confirm the first mover’s trust by sharing the fair half of the gains (cooperate), or to keep the full amount (defect). Figure 1 displays the players’ moves and the possible payoffs.

Figure 1. Trust game dilemma.

This research aimed at investigating which of two mechanisms (reputation or reputation with apology) was most effective for resolving and alleviating a trust breakdown (defection in the trust game). The offender (trustee) was a member with an unblemished historical track, who breached the norm (defected in the game) as a result of an error, experiencing and expressing regret thereafter. In exploring this particular scenario, it was important to retain control over the offender’s profile across the two treatments. Thus, trustees were simulated in this study and participants were always assigned to the trustor’s role.

We expected two main outcomes, the first related to forgiveness and the second concerned with the restoration of trust. Firstly, apology has been shown to predict forgiveness by evoking the victim’s empathy for the offender [4]. As such, participants who viewed the offender’s apology online were expected to report increased forgiveness towards him/her as compared to participants viewing the offender’s reputation score only (Hypothesis 1). In the second set of outcomes we expected that participants, who had viewed the offender’s apology and had presumably forgiven the offender, would once again trust the offender. Trust was indirectly measured by reputation and cooperation. In online auctioning, users trust members whose reputations are high [6]. This finding was inverted and applied to the case of trust restoration. If participants’ trust in the offender is repaired, they should give the offender higher reputation scores directly after the offence and also in any future interactions (Hypothesis 2). Furthermore, it has been shown that members whose trust in others is high are more likely to cooperate with them [1]. Therefore, participants whose trust in the offender has been restored should be more willing to cooperate with the offender given a future interaction (Hypothesis 3).

3.2 Participants and Procedure
An email announcement for participation was sent across four different departments of a UK university promising monetary reward. In total, forty-two participants responded and took part in this study. It was made clear in the announcement that proficiency in English was a requirement for attending. All participants were enrolled in undergraduate studies and were between 18 and 24 years of age.

Participants arrived in three groups of twenty-one. They were then taken to a quiet room and each one of them seated by a computer. None of the participants faced one another and it was ensured they could not see each others’ screens. The trust game was played using a web-based application. In making the game appear believable, participants were led to believe they were assigned to either first movers (trustor) or second movers (trustee) by a random draw. However, participants were always first movers and faced the dilemma of trusting the second mover by giving the points or withdrawing from the risk involved by keeping the 50 points. The second mover was always simulated in each round with a new non-gender indicative pseudonym. Participants were instructed that in each round they were playing with a new remote player from the University of Geneva. To avoid end-game effects (i.e. decreased trust in the final rounds of the game), usually observed in these kinds of games (e.g. [1]), participants were told they would randomly play up to 10 rounds of the game. After each round, an onscreen notification informed them whether they would proceed to the next round. In the experiment, the rounds played were set to six rounds in total.

A “starred” reputation system was embedded in the trust game application; one star signalled the lowest reputation score obtainable and five stars the highest. The stars summarise the mean of reputation scores received in previous transactions, similar to the Amazon auctioning reputation system [6]. Second movers’ reputation scores in each round were simulated and held constant for all participants; however participants were instructed that reputation scores were calculated on the basis of other players’ ratings and were constantly updated. The same reputation mechanism was made available to participants for rating the second mover, whenever they made the initial trust move.

In the first four rounds of the game participants played with a different simulated second mover. To make the second movers believable, each one displayed a different reputation score ranging from two, the lowest, to five, the highest; the second mover of round one had no reputation score displayed since he had not been previously rated. During the first four rounds, if trusted, the second movers always reciprocated the initial trust move by returning the fair amount. In round 5, the second mover for the first time betrayed the participant’s trust by returning nothing. At this stage, participants received one of two treatments: reputation or reputation with apology (see Section 3.3). At the end of round 5, participants were asked to assign a reputation score to the second mover. Immediately after, they filled out a questionnaire measuring forgiveness. At the beginning of round 6, participants were informed they were randomly assigned to the same player again, i.e. to the offender of round 5. The reputation score given to the offender in round
4.1 Validation of forgiveness measure

A principle component analysis of the modified twelve-point forgiveness questionnaire showed two factors with Eigenvalues greater than 1.0, presumably one for revenge and one for avoidance. An examination of the communalities table (Varimax rotation) suggested that four items be omitted as they loaded on both factors (>0.4). The remaining items all met the threshold criteria of 0.4 on one of the two factors. The final eight-point questionnaire consisted of four questions measuring revenge and four questions measuring avoidance (Appendix A).

4.2 Analysis

A multivariate analysis was performed with condition as a factor (reputation or apology) on the following measures: reputation score after round 5, reputation score after round 6, cooperation and forgiveness. Table 1 summarises the results.

<table>
<thead>
<tr>
<th>Condition</th>
<th>N=20</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation</td>
<td>Reputation</td>
<td>0.70</td>
<td>0.47</td>
</tr>
<tr>
<td>Apology</td>
<td>1.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Forgiveness</td>
<td>Reputation</td>
<td>25.65</td>
<td>7.47</td>
</tr>
<tr>
<td>Apology</td>
<td>23.10</td>
<td>7.39</td>
<td></td>
</tr>
<tr>
<td>Reputation score after round 5</td>
<td>Rejection</td>
<td>1.90</td>
<td>1.68</td>
</tr>
<tr>
<td>Apology</td>
<td>2.95</td>
<td>2.08</td>
<td></td>
</tr>
<tr>
<td>Reputation score after round 6</td>
<td>Rejection</td>
<td>4.20</td>
<td>3.20</td>
</tr>
<tr>
<td>Apology</td>
<td>6.70</td>
<td>2.20</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Summary of results.

The main effect of condition (reputation, apology) was non-significant (F (4, 35) = 2.48, p=0.05). After round 5, participants completed a measure of forgiveness. Participants in the apology condition as compared to those in the reputation condition did not report more forgiveness (F (1, 38) = 1.17, p>0.05). The first hypothesis was not supported. After the offender’s defection in round 5, participants in the apology condition gave reputation scores to the offender that were not significantly higher than the scores given by participants in the reputation condition. However, the trend was in the predicted direction (F (1, 38) = 3.06, p=0.09). In round 6, once the offender had shared the fair half of the gains, participants of the apology condition (F (1, 38) = 8.26, p<0.01) condition assigned significantly higher reputation scores to the offender as compared to participants of the reputation condition. The second hypothesis was partly confirmed; participants who received an apology assigned higher reputation scores but only after round 6 when the offender had cooperated. Finally, in support of the third hypothesis, in round 6, participants in the apology condition made the initial trust move by transferring the 50 points more frequently than participants in the reputation condition (F (1, 38) = 8.14, p<0.01).

5. DISCUSSION

Participants, who viewed the offender’s reputation score only, compared to those who also received an apology, demonstrated lower trusting behaviours towards the offender. The offender’s displayed reputation score in the round previous to the offence was the highest obtainable, thus signalling the offender’s prior benevolence to others. This, in combination with the offender’s gesture to return the fair amount in round 6 did not express the offender’s low intention when defecting and high intention when repairing. As a result, participants’ trust in the offender was not restored. This can be seen by the low reputation scores participants assigned after the offender’s return of the fair amount in round 6. Even more, participants of the reputation condition frequently withheld the initial trust move in the round after the offence. The risk involved in the trust move was only
25 pence. Hence, it appears plausible that some participants withdrew from the transaction as a consequence of retaliation rather than only due to a lack of trust. These conclusions taken together suggest that a victim’s trust in an unintentional offender cannot be restored with a reputation system only, while misattributions may be made that lead to retribution.

This study further compared the reputation system to a system that also displayed the offender’s apology to the victim. We anticipated that the latter, would elicit higher degrees of forgiveness and an increase in trust for the offender. The results found partly confirmed this prediction. When viewing an apology from the offender directly after the offence, participants did not immediately report increased forgiveness. Similarly, participants’ trust towards the offender was not restored; this can be seen in the low reputation scores they assigned. By contrast, in round 6, the offender’s apology, as compared to viewing the offender’s impeccable reputation score only, motivated victims to always cooperate with the initial trust move. After round 6, when the offender had corroborated his/her apology with the reparative move, participants in the apology condition assigned higher reputation scores than those who had not viewed the apology. Given the apparent restoration of trust in round 6, we postulate that participants also forgave the offender; retrospectively, the forgiveness measure should have been included also after round 6. Nonetheless, this research does not give a clear answer on why the offender’s apology did not repair the offence directly after it occurred in round 5. Was it due to the lack of a prior history between the two parties? Or was it because of the lack of non-verbal cues online, e.g. blush, which make it more difficult to discriminate between a truthful account and one that is deceitful? In summary, it appears that online apologies can restore trust but only when the offender has acted on his/her words of regret. In this sense, the eBay “mutual feedback withdrawal” forum prompts both parties to sustain the interaction, thus allowing the offender to repair his or her actions which in turn can restore the victim’s trust. However, this approach does not shield both parties from the initial dismay resulting from the offence. On the one hand, the victim may feel deceived and thus experience anger towards the offender. Conversely, the offender may also experience anger due to the unjust low rating. In a panel of users collected before the introduction of the “mutual feedback withdrawal” feature, eBay sellers who received negative feedback were 25% less likely to post new products than sellers with only positive feedback; this effect was attributed to users’ feelings of injustice when receiving an unfair punishment [4]. This short paper does not undertake the task to propose specific solutions that address this issue. It is important though to note that [10] have investigated ways for proactively facilitating forgiveness in social systems.

To conclude, this work set its focus on understanding whether reputation systems are able to repair trust as well as they sustain it and furthermore, whether apologies as given online are sufficient to motivate repair. As expected, it was found that reputation systems cannot repair trust breakdowns. By contrast, an apology from the offender restored the victim’s trust, although only after the offender had proven his/her intent by correcting his/her future behaviour. We end this discussion with some open questions for social system designers. Is it always feasible for a benevolent member who has unintentionally transgressed a norm to repair his/her actions? If not, should systems take upon themselves the role to resolve the trust breakdown? Furthermore, should social systems allow for the unfair punishment of a benevolent member or should the system safeguard such a member in the first place?

6. APPENDIX A

“I wish that something bad will happen to him/her in the game”, “I want him/her to get what he/she deserves in the game”, “If I get the opportunity in the game, I’m going to get even”, “I would like to see him/her hurt and miserable in the game”, “If I play with her/him again, I will pretend like he/she doesn’t exist, isn’t around”, “If I play with her/him again, I will avoid him/her”, “If I play with her/him again, I will cut off the relationship”, “If I play with her/him again, I will withdraw from him/her”.

7. ACKNOWLEDGMENTS

Jens Riegelsberger (Google) is acknowledged for his comments on an earlier draft of this paper. We thank Anjum Hanif (eBay) for his advice on the eBay system. We further acknowledge Arvind Bhusate, Tom Carlson and Daniel Ramirez who assisted in the experiment. This research was funded by the HUMAINE Network of Excellence.

8. REFERENCES


