Available online at www.sciencedirect.com
ScienceDirect
Int. J. Human-Computer Studies 66 (2008) 466–480
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

In praise of forgiveness: Ways for repairing trust breakdowns in one-off online interactions

Asimina Vasalou[a,*], Astrid Hopfensitz[b], Jeremy V. Pitt[a]

[a]*Electrical and Electronic Engineering, Imperial College London, Exhibition Road, SW7 2BT London, UK*
[b]*Toulouse School of Economics (GREMAQ), Manufacture des Tabacs, 21 Allée de Brienne, 31000 Toulouse, France*

Received 13 June 2007; received in revised form 22 January 2008; accepted 12 February 2008
Communicated by S. Wiedenbeck
Available online 10 March 2008

Abstract

Online offences are generally considered as frequent and intentional acts performed by a member with the aim to deceive others. However, an offence may also be unintentional or exceptional, performed by a benevolent member of the community. This article examines whether a victim's decrease in trust towards an unintentional or occasional offender can be repaired in an online setting, by designing and evaluating systems to support forgiveness. We study which of three systems enable the victim of a trust breakdown to fairly assess this kind of offender. The three systems are: (1) a reputation system, (2) a reputation system with a built-in apology forum that may display the offender's apology to the victim and (3) a reputation system with a built-in apology forum that also includes a "forgiveness" component. The "forgiveness" component presents the victim with information that demonstrates the offender's trustworthiness as judged by the system. We experimentally observe that systems (2) and (3), endorsing apology and supporting forgiveness, allow victims to recover their trust after online offences. An apology from the offender restores the victim's trust only if the offender cooperates in a future interaction; it does not alleviate the trust breakdown immediately after it occurs. By contrast, the "forgiveness" component restores the victim's trust directly after the offence and in a subsequent interaction. The applicability of these findings for extending reputation systems is discussed.

Keywords: Trust; Social dilemmas; Forgiveness; Reputation; Apology; One-off interactions

1. Introduction

Trust is a social lubricant for computer-mediated communication (CMC) enhancing collaboration, cooperation and information exchange (Rocco et al., 2000; Bos et al., 2002; Ridings et al., 2002). It is thus not surprising that a large body of trust research has focussed on the factors that engender trust. For instance, in online auctions, the effect of sellers' reputation on buyers' trust, as manifested in their bidding offers, has been repeatedly investigated (e.g. Ba and Pavlou, 2002; Lucking-Reiley et al., 2007; Resnick et al., 2006). This body of work has been important as it has offered prescriptions for the design of trust-enabling social systems. Despite the value of this approach though, the current debate has failed to fully address the need for repair mechanisms, offered when trust *breaks down*.

This oversight can be partly attributed to a research focus on intentional and frequent acts performed by an offender with the aim to deceive others. For example, impersonators often become part of close-knit communities under a contrived identity. When discovered, identity deception can damage the trust cultivated within the wider community as users begin to question each others' motives (Grady, 1998; Joinson and Dietz-Uhler, 2002). In addition, "trolls", who are seemingly sincere users, seek simply to provoke a reaction by posting contentious comments which challenge a community's commonly held beliefs (Donath, 1998). To eradicate such intentional and recurrent offences, social system designers have created tools

*Corresponding author. Tel.: +44 7796143213; fax: +44 2075946274.
E-mail addresses: a.vasalou@imperial.ac.uk (A. Vasalou), astrid.hopfensitz@univ-tlse1.fr (A. Hopfensitz), j.pitt@imperial.ac.uk (J.V. Pitt).

1071-5819/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2008.02.001

that allow communities to self-regulate: in newsgroups, malignant members can be filtered out, their comments ceasing to affect others; in online markets, a fraudulent seller might be assigned a low reputation rating, thereby being driven out of the market. If the behaviour escalates, explicit measures can be taken by appointed individuals from the community stakeholders (e.g. in newsgroups intervention by a moderator; in Wiki, an administrator restoring original pages). In the last resort, there is also recourse to legal action.

As opposed to this dominant interest in intentional breakdowns, some researchers have alluded to certain situations during which one party may violate another's trust *unintentionally* (e.g. Riegelsberger et al., 2005a). For instance, an expected action may not be executed because the trusted party is unable to perform or does not have the required ability to do that action: as an example, in a remote collaboration project, trust was endangered as senior members of one remote team made false assumptions about their junior counterparts' skills (Rocco et al., 2000). Moreover, the fulfilment of obligations may be prevented by unforeseen events that are outside the trusted party's control. A promised delivery might be delayed due to slow postal services (Riegelsberger et al., 2005a). Online markets recognise these possibilities, and buyers are advised to be patient, reverting to negative feedback only when necessary (Kollock, 1999); despite these words of caution, eBay sellers still report distressing experiences due to unwarranted or disproportionate negative feedback (Khopkar et al., 2005). Also, when placed in a social dilemma under conditions of anonymity, as it happens for instance in an e-commerce environment, members are more likely to "defect", i.e. not cooperate, and by doing so to damage the well-being of others (Bos et al., 2002; Zheng et al., 2002; Vasalou et al., 2006a). However, although a member may "slip" once, it has been shown that when alerted, most will repair the damage by apologising and correcting their future behaviour (Vasalou et al., 2006a).

The above scenarios concern a different kind of offender: a benevolent member of the community who may have breached a norm unintentionally, or slipped once, regretting his/her behaviour thereafter. In face-to-face communication, proximity allows the offender to apologise, to elaborate on his/her intentions and to repair the breakdown, thus paving the way towards forgiveness (Boon and Sulsky, 1997; McCullough et al., 1998). Moreover, non-verbal expressions (e.g. blushing) given off by the offender may supersede and complement words of regret (Castelfranchi and Poggi, 1990; Miller, 1996; Keltner and Buswell, 1997). Anonymous, one-off encounters which are encouraged in many online settings add complexity in resolving such offences. This is partly due to the narrow time window of each interaction in combination with the impoverished communication channel constricting the cues of trustworthiness one can acquire about another member (e.g. integrity, willingness to comply to institutions, benevolence; Riegelsberger et al., 2005a). At the onset, this establishes interactions that are perceived as more risky, thus building emotional barriers that may stand in the way of resolution if trust breaks down. Within this high-risk interaction, reputation systems, which operate to maintain trust, are not effectively designed to repair trust. In our view, reputation systems have taken little provision to encourage the repair of trust breakdowns; we are aware of only two approaches that have been recently proposed or implemented in this context.[1]

A first approach has been introduced in eBay's electronic marketplace: using a feature called the "mutual feedback withdrawal" users can retrospectively contest the reputation score they received. Only if both the victim and offender agree to engage in this process, then the resolution of the issue is taken offline. Later, the victim of the offence 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. In a second approach, Vasalou and Pitt (2005) and Vasalou et al. (2006b) have proposed to facilitate resolution by embedding an intelligent "forgiveness" component within reputation systems. In their conceptualisation, upon a breakdown in trust, the system will detect the trustworthiness of the offender by taking into account a number of factors. Only if the offender is judged positively by the system, the victim is presented with those factors to consider before assigning a reputation score to the offender. Essentially, this intervention intends to alleviate the victim's possibly negative attributions (e.g. the offender intended to perform the offence) while at the same time it aims to shield the unintentional/infrequent offender from receiving an unjust punishment, i.e. in the form of a low rating. However, it is as yet unclear whether either of these approaches, as compared to a basic reputation system, is useful for facilitating repair after an unintentional act of an infrequent offender, who happens to be a valued member of the community. This article sets out to answer this question by comparing an "apology" and a "forgiveness" approach against a reputation system.

Before we begin to discuss how to repair trust, Section 2 defines online trust. We go on to show how trust is sustained during one-off interactions with the use of reputation systems. An example is given to demonstrate the harmful consequences of reputation mechanisms that function without the operation of a reparative mechanism.

[1]Though repairing trust breakdowns in one-off mediated interactions have not received adequate attention, we acknowledge two related efforts in the wider field of human–computer interaction. In ambient environments agents pursue their users' interests in an autonomous manner. Conflicts that inevitably emerge can be reduced by integrating forgiveness and regret into the trust framework that governs the agents' actions (Briggs and Marsh, 2006). In using computer applications, users often encounter error messages which can temporarily obstruct their work. Apologetic messages have been tried out as a way to alleviate users' frustration during these system errors (Tzeng, 2004).