Contents lists available at SciVerse ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

Building and rebuilding trust with promises and apologies[☆]

Eric Schniter^{a,*}, Roman M. Sheremeta^b, Daniel Sznycer^c

^a Economic Science Institute, Chapman University, One University Drive, Orange, CA 92866, United States

^b Argyros School of Business and Economics, Chapman University, One University Drive, Orange, CA 92866, United States

^c Center for Evolutionary Psychology, University of California, Santa Barbara, CA 93106, United States

ARTICLE INFO

Article history:

Received 14 May 2012

Accepted 8 September 2012

Available online xxx

Keywords:

Promise

Atonement

Apology

Cheap talk

Cheap signals

Trust game

Trust building

Remedial strategies

Reciprocity

Experiments

ABSTRACT

Using trust games, we study how promises and messages are used to build new trust where it did not previously exist and to rebuild damaged trust. In these games, trustees made non-binding promises of investment-contingent returns, then investors decided whether to invest, and finally trustees decided how much to return. After an unexpected second game was announced, but before it commenced, trustees could send a one-way message. This design allowed us to observe the endogenous emergence and natural distribution of trust-relevant behaviors and focus on naturally occurring remedial strategies used by promise-breakers and distrusted trustees, their effects on investors, and subsequent outcomes. In the first game 16.6% of trustees were distrusted and 18.8% of trusted trustees broke promises. Trustees distrusted in the first game used long messages and promises closer to equal splits to encourage trust in the second game. To restore damaged trust, promise-breakers used apologies and upgraded promises. On average, investments in each game paid off for investors and trustees, suggesting that effective use of cheap signals fosters profitable trust-based exchange in these economies.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

In modern economies, where trust realizes vast amounts of potential gains in transactions involving deferred or risky returns, problems associated with developing and restoring trust are particularly relevant. A scientific understanding of the processes that encourage trust where it did not previously exist and restore trust when it is damaged is therefore of paramount importance. Despite the large literature on damages to corporate reputation (e.g. see Barnett, 2003 on US chemical industry disasters; see Robinson and Rousseau, 1994 for a survey of corporate trust violations), very little research exists on how new trust can be encouraged where it did not previously exist and how damaged trust can be rebuilt (Dirks et al., 2009). Most of the existing research in this area (but see Fischbacher and Utikal, 2010) is either exclusively theoretical (Lewicki and Bunker, 1996; Mishra, 1996; Lewicki and Wiethoff, 2000; Ren and Gray, 2009; Gillespie and Dietz, 2009), based on anecdotal or archival evidence (Elsbach, 1994; Knight and Pretty, 1999), surveys (Slovic, 1993), diary studies (Conway and Briner, 2002), fictional vignettes (Tomlinson et al., 2004), videotaped dramatizations (Kim et al., 2004, 2006),

[☆] For inspiration to pursue this study we thank John Dickhaut. We thank an advisory editor and an anonymous referee for their comments. Helpful comments were also received from Hilly Kaplan, Wojtek Przepiorka, and participants at the Workshop on Communication in Games (at the University of Zurich), the Human Behavior and Evolution Society annual meeting (in Montpellier, France), the Center for Evolutionary Psychology (at UC Santa Barbara), the John Dickhaut Memorial Conference (at Chapman University), and the Association for Religion, Economics and Culture annual meeting (at Chapman University). We would also like to thank the Economic Science Institute at Chapman University for funding this research.

* Corresponding author.

E-mail address: eschniter@gmail.com (E. Schniter).

or experimental designs using deception (Gibson et al., 1999; Bottom et al., 2002; Nakayachi and Watabe, 2005; Schweitzer et al., 2006; Ohtsubo and Watanabe, 2009).

To study how damaged trust can be rebuilt and new trust can be encouraged, we conducted a non-deceptive study wherein financially motivated participants used endogenously created and naturally distributed promises and apologies. Our study is based on a version of the “investment game” by Berg et al. (1995). In the original investment game an investor is endowed with \$10 and can invest any portion of her endowment by sending it to a trustee. The amount sent triples in value before reaching the trustee. Having received funds from this tripled investment, the trustee can reciprocate by returning any portion of these funds to the investor. Since sending money is risky, investments are usually interpreted as trust, and since returning money is costly, reciprocation via returns on investments is interpreted as evidence of trustworthiness.¹ The investment game, therefore, has been extensively used to study trust and reciprocity in an investment setting (for a review see Ostrom and Walker, 2005). A common finding in the literature is that investors tend to exhibit trust and trustees tend to reciprocate. It has also been well established that pre-play communication, even if “irrelevant” to game strategy, can induce higher contributions in public goods games (for meta-analyses see Sally, 1995; Balliet, 2010) and more cooperation in dyadic social dilemmas (Deutsch, 1958, 1960; Radlow and Weidner, 1966; Buchan et al., 2006; Duffy and Feltovich, 2006; Bracht and Feltovich, 2009). However, with the exception of a few studies using deception, the experimental economic literature is silent as to what behavior ensues when promises fail to establish trust and what happens to trust and reciprocity in subsequent interactions after promises are broken and trust is damaged.

In this paper we describe a study using trust games that examines how promises and messages are used to build new trust where it did not previously exist and to rebuild damaged trust. In these games, trustees made non-binding promises of investment-contingent returns, then investors decided whether to invest, and finally trustees decided how much to return. After an unexpected second game was announced, but before it commenced, trustees could send a one-way message. This design allowed us to observe the endogenous emergence and natural distribution of trust-relevant behaviors and focus on naturally occurring remedial strategies used by promise-breakers and distrusted trustees, their effects on investors, and subsequent outcomes. In the first game 16.6% of trustees were distrusted and 18.8% of trusted trustees broke promises. Trustees distrusted in the first game used promises closer to equal splits and – compared to previously trusted promise-keepers – relatively longer messages to encourage new trust in the second game. Promise-breakers used relatively higher new promises (compared to all other trustees) and messages (usually with apology) to successfully restore damaged trust. On average, investments in each game paid off for investors and trustees, suggesting that the context-specific signaling described above, can foster profitable trust-based exchanges in these economies.

2. Background

While mutually beneficial non-binding agreements help realize opportunities to gain from asynchronous trade, they are subject to exploitation by under-reciprocators or non-reciprocators. Our research focuses on trustees' cue and signal effects on investor trust in asynchronous exchanges that provide opportunity for mutual advantage. In these exchanges, we define trust as voluntarily ceding resources to another in the expectation that the other intends to reciprocate in accordance with signaled intentions. Trustworthiness is defined as reciprocation (of resources ceded by the investor) in accordance with signaled intentions.

To successfully navigate a trust-based cooperative interaction and avoid exploitation by cheaters, it is important for investors to obtain accurate information about the ability and willingness (propensity) of trustees to carry out their end of the cooperative deal. Trustworthy reputations that have been demonstrated by past actions serve as reliable *cues* upon which investors can make trust-based decisions. In the initial interactions with unknown partners, informative cues about an investor's willingness to trust or a trustee's trustworthiness are unavailable. In the absence of information about the interactants' past behavior, *signals*² are often sent to receivers with the intention to communicate information about the sender (e.g. see Farrell and Rabin, 1996); for example, that the sender is trustworthy. Where cues have informed investors of untrustworthiness, signals may be sent with the intention of persuading those investors that the sender is more trustworthy than inferred from those cues alone.

Signals encouraging trust appear to be important tools for developing mutually beneficial relationships under conditions where trust has not yet been established and where trust has been damaged. Without the effective use of signals cooperative

¹ This interpretation is based on the assumption that participants identify and act in accordance with unstated *if-then* propositions and expect others to as well (Rousseau, 1989), though there is no contract stating expected or contingent behavior in the classic “investment game” (see Berg et al., 1995). Because the assertion that the original game was universally understood to be about “trust” was debatable, John Dickhaut preferred calling it the “investment game” – as it is in the Berg et al. (1995) article. By adding a new starting stage to the game where trustees make promises to return a portion of income from investment – this game becomes a game more explicitly about trust. For this reason we refer to our modified form of the classic investment game, described below, as a “trust game.”

² We distinguish *cues* from *signals* from *coercion* (borrowing from similar definitions by Diggle et al., 2007; Scott-Phillips, 2008) as follows. *Cue*. Any behavior or feature that (i) affects the behavior of other organisms; (ii) which is effective because the effect has evolved to be affected by the behavior or feature; but which (iii) did not evolve. *Signal*. Any behavior or feature that (i) affects the behavior of other organisms; (ii) evolved because of those effects; and (iii) which is effective because the effect (the response) has evolved to be affected by the behavior or feature. *Coercion*. Any behavior or feature that (i) affects the behavior of other organisms; (ii) evolved because of those effects; but which (iii) is effective for some reason other than that the effect has evolved to be affected by the behavior or feature.

interactions may be foregone: potential investors may decide not to extend trust when they lack reputational information and when cues indicate a breach of trust. Further, when trust has been damaged, signals give investors access to relevant though invisible propensities of trustees, such as in the case of recalibrated upgrades in trustworthiness. This is true whether trust has been damaged intentionally or unintentionally (Axelrod and Dion, 1988).

Although signals that accurately convey behavioral propensities are potentially useful to both senders and receivers, signalers may send “dishonest signals” to benefit at the expense of receivers. Critical receivers can incur lower costs than naïve receivers (Dawkins and Krebs, 1978; Maynard-Smith, 1982), and so natural selection favors those receivers who can accurately assess the cost–benefit tradeoffs associated with emitters’ signals, and calibrate their trustfulness accordingly.

Zahavi (1975) addressed the question of “why are signals reliable?” suggesting that the high production cost of a signal guarantees its reliability, insofar as the production cost outweighs the benefits gained from using the signal deceptively, but not from using it honestly. The prototypical example is the massive and colorful peacock’s tail, indexing the peacock’s genetic quality for peahens’ mate selection (Petrie et al., 1991). In this system, costly signals persuade the receivers while cheap signals fail to do so (Zahavi, 1993; Grafen, 1990).

Production costs are not the only warrantors of signal reliability, however. Human language, whether spoken or written, is an arbitrary communication system that often uses relatively cheap-to-produce signals to negotiate trust between individuals with conflicting interests (Lachmann et al., 2001). If these cheap signals are used by and relied upon by humans globally and on average, what explains the maintenance of their reliability?

The reliability of cheap signals is supported by the actuality or threat of social sanctions that can more than offset the short-term benefits of cheating and deception (Rohwer, 1977; Kiyonari et al., 2000; Masclet et al., 2003). A selective regime characterized by repeated interactions among known others (Kelly, 1995) has led to psychological mechanisms for social exchange that balance (i) the costs of mistaking a one-shot interaction for a repeated interaction with (ii) the far greater costs of mistaking a repeated interaction for a one-shot interaction (Delton et al., 2011). Hence, participants in explicitly one-shot anonymous experiments often behave as if they expect repeated interactions with trustworthy, intrinsically valuable partners (e.g. see Hoffman et al., 1996; Kiyonari et al., 2000).

While the sanctioning of false signals and our tendency to err to caution may reduce the frequency of false signals in a population, those individuals who expect to escape sanctions may be more motivated to use signals deceptively. In economies where opportunity costs of forgone trust-based exchange are larger, receivers tend to tolerate greater proportions of false signals to honest signals. Specifically, the logic of error management theory (for a review see Haselton and Nettle, 2006) predicts that despite the existence of false signaling and the costs of receiving false signals, signals will tend to be received when opportunity costs associated with not receiving true signals of trustworthiness (from forgone advantageous exchange) are greater than costs associated with receiving false signals of trustworthiness (i.e. when the consequent exchange produces a loss). This economically justified tolerance of a rate of false signaling also predicts that individuals will exploit opportunities to use deception. This study explores the use of cheap signals (e.g. promises of reciprocation, personalized messages, and apologies) that do not directly affect payoffs of the game, or require monetary costs for production, yet are common features of trust-based interactions. Personalized communication may improve cooperation (Orbell et al., 1988; Bohnet and Frey, 1999; Ridings et al., 2002; Zheng et al., 2002; Buchan et al., 2006) by facilitating coordination, decreasing social distance, raising solidarity, and providing the cues of familiarity that are normally associated with trustworthy relationships. Non-binding promises have also been shown to increase cooperation (Rubin and Brown, 1975; Kerr and Kaufman-Gilliland, 1994; Elingsen and Johannesson, 2004; Charness and Dufwenberg, 2006). In relationships where trust has been damaged, apologies and explanations have been shown to elicit forgiveness (Ohbuchi et al., 1989; Tavuchis, 1991; Lewicki and Bunker, 1996; Benoit and Drew, 1997; Girard and Mullet, 1997; McCullough et al., 1997, 1998; Girard et al., 2002; Witvliet et al., 2002) and promote future trust (De Cremer et al., 2010). These strategies are based on signals that are cheap to produce, raising the questions of how people use signals in these contexts; when the signals achieve their intended effects; and who benefits from their use.

In sum, while cheap signals are helpful for building new trust and rebuilding damaged trust to achieve efficient outcomes, they can be used deceptively and may be distrusted, making their reliability tenuous. Therefore, we expect that in our experiment trustees whose actions have already produced reliable cues establishing their trustworthy reputations (by keeping promises and not succumbing to more profitable opportunism) will be less incentivized (than previously distrusted trustees, or trustees whose reputations indicate untrustworthiness) to spend time and effort constructing messages to persuade investors to trust them, when those messages might be distrusted. Previously distrusted trustees who have not established trustworthiness and untrustworthy trustees (i.e. promise-breakers) are expected to make use of promises and messages in an attempt to affect investors’ decisions to trust. We also expect that when used and “working” to affect investors’ trust, signals conveying a trustworthy propensity will provide benefits to both investor and trustee on average.

3. Experimental design and procedures

The experiment was conducted at Chapman University’s ESI laboratory. 458 participants (229 pairs) were recruited from a standard campus-wide subject pool for participation in an experiment that could last up to 45 min. Participants interacted with each other anonymously over a local computer network. The experiment, which lasted an average of 35 min total and did not involve deception, proceeded as follows. Upon arrival, participants in the experiment were told that they would receive \$7 for participation, to be paid at the end of the experiment. Participants then received instructions (see Appendix A)

for a single trust game with (i) no indication of a subsequent game to follow and (ii) no promises that the experiment would end at conclusion of that game.

Participants were assigned to one of two roles: “Participant A” (investor), or “Participant B” (trustee). First, the trustee completed the following standardized statement (which we will refer to below as a *promise*) by selecting a natural number amount from 0 to 20: “I (Participant B) promise to transfer back \$___ of my income to you (Participant A) if you choose IN”. This statement was not binding, however. That is, the trustee was *not* obligated to transfer back the amount promised to the investor, and both trustee and investor knew this. The computer conveyed the trustee’s statement to the investor and then the investor chose either OUT or IN. If the investor chose OUT, she received \$5 and the trustee \$0. If the investor chose IN, then the trustee received \$20 (the “income”), after which he selected a whole dollar amount from \$0 to \$20 to send back to the investor.

After the first trust game (Game 1) finished, participants were given instructions (see [Appendix A](#)) indicating that a second trust game (Game 2) would follow. In Game 2, participants were told they would remain in the same roles and interact with the same partner as in Game 1. However, prior to Game 2, the trustee was given an opportunity to use a text box to send a one-way message to the investor. Trustees were told that “in these messages, no one is allowed to identify him or herself by name, number, gender, or appearance,” but that other than these restrictions, trustees could “say anything in the message.” If trustees wished not to send a message they were instructed to “simply click on the send button without having typed anything in the message box.” The computer conveyed the trustee’s message and subsequently the standardized promise to the investor, and then Game 2 proceeded. We specified that Game 2, which had the same rules as Game 1, was the last and final part of the experiment (i.e. there would be no subsequent games).³

There were 25 experimental sessions. Each session had between 10 and 24 participants. The average experimental earnings were \$18, ranging from a \$0 to \$40, plus \$7 for arriving to the experiment on time and participating. No participant participated more than once, and no participant had prior experience with a similar game environment.

4. Results

4.1. Game 1

We expect that trustees, aware of investor self-interest and motives for critical signal reception, would promise investors transfers of at least \$6 (minimally higher than the payoff to the investor if he chooses OUT) but less than \$20 (which would provide no benefit to the promise-maker). Two plausible focal points for promised return amounts are the midpoint of the \$6–\$19 range, \$12.5 (though only whole dollar amounts like \$12 or \$13 could be chosen), and the even-split of \$10. Wary that trustees’ may have less incentive to honor promises closer to \$20 than to the even-split amount of \$10, we also expect that investors should be more suspicious of the veracity of higher promises and therefore less likely to invest in higher promises. If the mind errs to caution, as we have suggested, and interprets the one-shot game as potentially repeatable, trustees who have been trusted should reciprocate enough to, at minimum, provide investors profitable returns on their investments. These predictions stand in stark contrast to the rational (non-cooperative) choice predictions that expect non-binding promises to have no effect on investors. According to rational choice theory, trustees who receive incomes should return nothing (despite what they may have promised) and, based on this, investors should always choose to not invest (regardless of the promise they received).

[Fig. 1](#) displays the aggregate distribution of investment and promise-keeping decisions in the experiment, while [Fig. 2](#) displays the distribution of promises made by trustees in Game 1. In Game 1, trustees on average promised to return \$9.20 ($SD = 2.38$) out of \$20 and 83.4% (191/229) of investors chose IN.

First we evaluate the distributions of Game 1 promises associated with trusting and distrusting decisions, respectively, and how these investment decisions affected investor and trustee earnings. The distribution of promises in [Fig. 2](#) indicates that investors who chose IN received promises in the range of \$6–\$19 (99% of the time), with promises of \$12 or \$13 relatively uncommon (1% of the time), and the promise of \$10 most common (more than 50% of the time). Investors who chose OUT received lower promises on average (i.e. $M_1 = \$8.61$ ($SD_1 = 4.33$) versus $M_2 = \$9.31$ ($SD_2 = 1.75$); Wilcoxon rank-sum test, p -value = 0.01, $n_1 = 191$, $n_2 = 38$), and, compared to trusted promises, received either relatively low or relatively high promises overall. To confirm this observation, we estimate probit models (see [Table 1](#), specifications 1 and 2), where the dependent variable is the investment decision in Game 1 and the independent variables are dummy variables for promises less than \$9 and greater than \$11, as well as the amounts of these promises. In specification (1), the dummy variables are negative and significant, indicating that investors are less likely to invest when promises are either relatively low or relatively high. Moreover, specification (2) indicates that, among promises lower than \$9, there is a positive correlation between the amount promised and the probability of investment. On the other hand, among promises higher than \$11, there is a negative correlation between the amount promised and the probability of investment. In other words, promises closer to the even-split of \$10 elicit a higher rate of IN responses.

³ After each trust game participants were also asked to fill out a 20 item survey in which they reported their emotional states consequent on their decisions, game interactions, and resulting outcomes. Analysis and discussion of the mediating roles of emotions are not included in this paper.

Fig. 1. Aggregate distribution of decisions in Games 1 and 2.

In Game 1, investment yielded greater payoffs than non-investment for both investors and trustees. Investors who chose IN received back \$8.19 on average, which is more than their original endowment of \$5 (Wilcoxon signed rank test, p -value < 0.01 , $n = 191$). Trusted trustees earned an average of \$11.81, more than the \$0 of distrusted trustees (Wilcoxon rank-sum test, p -value < 0.01 , $n_1 = 38$, $n_2 = 191$). The OLS estimation of specifications (3) and (4) in Table 1 indicates that the amount returned by trustee is correlated with the amount promised. Specifically, specification (3) indicates that both low and high promises are followed by lower returned amounts. Moreover, specification (4) indicates that amount returned increases as it gets closer to the even-split promise of \$10. These results support the rationale for why investors receiving especially high or low promises should tend to choose OUT.

For the investors who chose IN, the mean amount returned of \$8.19 was significantly lower than the mean trusted promise of \$9.31 (Wilcoxon signed rank test, p -value < 0.01 , $n_1 = n_2 = 191$). Despite mean returns being lower than promised, we find that promises tended to be veridical: 81.2% of trusted promises (155/191) were kept (i.e. the amount returned was equal to or greater than the promise), and 18.8% (36/191) were broken (i.e. the amount returned was less than the promise). Below we use the terms “promise-keepers” and “promise-breakers” to refer to trusted trustees who exactly met or exceeded their

Fig. 2. Distribution of promises in Game 1 (resulting in IN or OUT).

Table 1

Game 1 determinants of investment and return.

Dependent variable Specification	Invest1 (1) Probit	Invest1 (2) Probit	Return1 (3) OLS	Return1 (4) OLS
<i>Promise1</i> < 9 [dummy for promises less than 9 in Game1]	−0.83*** (0.21)	−4.81*** (1.07)	−1.77*** (0.41)	−6.92** (2.77)
<i>Promise1</i> < 9 × <i>Promise1</i> [amount promised in Game1 for promises < 9]		0.60*** (0.16)		0.72* (0.38)
<i>Promise1</i> > 11 [dummy for promises greater than 11 in Game1]	−1.01** (0.41)	8.44 (5.19)	−4.52*** (0.95)	12.30* (6.62)
<i>Promise1</i> > 11 × <i>Promise1</i> [amount promised in Game1 for promises > 11]		−0.55* (0.30)		−1.11** (0.43)
Constant	1.35*** (0.15)	1.35*** (0.15)	8.81*** (0.21)	8.81*** (0.21)
Observations	229	229	191	191

Standard errors in parenthesis.

* Significant at 10%.

** Significant at 5%.

*** Significant at 1%.

promised amounts, and who returned less than their promised amounts (whether the returns were monetarily profitable to the investors or not), respectively.

4.2. Game 2

While cheap signals are manipulated by trustees, affect investors, and provide net benefits to both investors and trustees in Game 1, facilitating profitable trust-based exchanges where previous reputations had not been established, Game 2 provides us a relatively different game environment in which to study cheap signals. In Game 2, reputations have been established for many investors and trustees – raising the question of whether the use of cheap signals will still matter where cues of trusting and trustworthy behavior (or its absence) are available.

In Game 2, trustees promised to return \$9.79 on average, a higher amount than the mean of \$9.20 promised in Game 1 (Wilcoxon signed rank test, p -value < 0.01, $n = 229$). Game 2 promises resulted in 87.3% (200/229) of investors choosing IN, only slightly more than the 83.4% (191/229) of IN decisions made in Game 1 (Fisher's exact test, p -value = 0.59, $n = 229$). Trustee reputation as established in Game 1 and the new promises issued in Game 2 affected investment decisions in Game 2. The estimation of probit models (specifications 1 and 2 in Table 2)⁴ indicates that both promise-breaking in Game 1 and promises lower than the even-split in Game 2 elicited less investment decisions in Game 2. Overall, the investments made in Game 2 paid off. Investors who chose IN received back \$8.73 on average, which is more than the OUT payoff of \$5 (Wilcoxon signed rank test, p -value < 0.01, $n = 200$). Trusted trustees earned an average of \$11.27, more than the \$0 of distrusted trustees (Wilcoxon rank-sum test, p -value < 0.01, $n_1 = 29$, $n_2 = 200$). The estimation of specifications (3) and (4) in Table 2 indicates that higher promises in Game 1, a greater extent of under-return relative to promise in Game 1, and uneven split promises in Game 2 all predict lower amounts returned in Game 2. Overall, and similar to Game 1, promises in Game 2 tended to be veridical; 75% of promises (150/200) were kept or exceeded, and 25.0% (50/200) were broken. In the sections below we further explore the effects of promises and messages on Game 2 investments and earnings across different “types” of dyads aggregated by Game 1 decisions.

4.2.1. Promise-keepers

For the subset of 155 Game 1 promise-keeping trustees, we observe higher average promises in Game 2. Fig. 3 displays the distribution of Game 2 promises resulting in IN and OUT made by these promise-keepers. Overall, this set of trustees promised to return an average of \$9.46 in Game 2, which is higher than their average promise of \$9.02 in Game 1 (Wilcoxon signed rank test, p -value < 0.01, $n_1 = n_2 = 155$).

Perhaps as a consequence of promise-keepers' demonstrated trustworthiness in Game 1, the Game 2 investment rate of 92.3% (143/155) in Game 1 promise-keepers was higher than the overall investment rate of 83.4% (191/229) in Game 1 (Fisher's exact test, p -value < 0.01, $n_1 = 155$, $n_2 = 229$). Specifications (1) and (2) in Table 3 indicate that messages with content and promises closer to the even split of \$10 positively affected investment in promise-keepers.⁵

⁴ In estimation of Table 2, we have excluded variables corresponding to the amount promised in Game 1 (i.e. *Promise1* < 9 × *Promise1* and *Promise1* > 11 × *Promise1*), since these variables are highly correlated with returns. Nevertheless, even when these variables are included the estimates in Table 2 are very similar.

⁵ Note that in estimating these regressions we cannot include both returns in Game 1 and promises in Game 1 since for promise-keepers they are perfectly correlated. Moreover, we had to omit variable *Promise2* > 11 × *Promise2* since there are only three observations greater than \$11, which makes *Promise2* > 11 and *Promise2* > 11 × *Promise2* almost perfectly correlated.

Table 2

Game 2 determinants of investment and return.

Dependent variable Specification	Invest2 (1) Probit	Invest2 (2) Probit	Return2 (3) OLS	Return2 (4) OLS
<i>Promise1</i> < 9 [dummy for promises less than 9 in Game1]	−0.05 (0.36)	−0.07 (0.36)	−0.85 (0.79)	−0.97 (0.78)
<i>Promise1</i> > 11 [dummy for promises greater than 11 in Game1]	−0.37 (0.65)	−0.21 (0.70)	2.67* (1.45)	1.73 (1.47)
<i>Return1</i> [return in Game1]	0.00 (0.12)	−0.02 (0.12)	0.02 (0.26)	0.03 (0.26)
<i>Broken1</i> [broken promise in Game1]	−0.80* (0.48)	−0.84* (0.49)	0.86 (1.33)	1.86 (1.36)
<i>Broken1</i> × (<i>Promise1</i> − <i>Return1</i>) [extent of under-return relative to Game1 promise]	−0.07 (0.11)	−0.10 (0.12)	−0.58* (0.30)	−0.64** (0.30)
<i>Distrusted1</i> [not trusted in Game1]	−0.22 (1.06)	−0.32 (1.10)	−1.53 (2.39)	−1.35 (2.38)
<i>Promise2</i> < 9 [dummy for promises less than 9 in Game2]	−1.37*** (0.32)	−2.13** (0.88)	−1.66** (0.74)	−1.47 (2.79)
<i>Promise2</i> < 9 × <i>Promise2</i> [amount promised in Game2 for promises < 9]		0.11 (0.12)		−0.01 (0.38)
<i>Promise2</i> > 11 [dummy for promises greater than 11 in Game2]	−0.14 (0.42)	2.20 (1.90)	0.78 (1.05)	−12.01** (4.64)
<i>Promise2</i> > 11 × <i>Promise2</i> [amount promised in Game1 for promises > 11]		−0.14 (0.11)		0.80*** (0.28)
<i>Constant</i>	2.05* (1.13)	2.26* (1.17)	8.98*** (2.51)	8.83*** (2.48)
Observations	229	229	200	200

Standard errors in parenthesis.

* Significant at 10%.

** Significant at 5%.

*** Significant at 1%.

We expected that, due to the greater reliability of available behavioral cues which demonstrated their trustworthiness (relative to the reliability of a cheap signal), Game 1 promise-keepers would be less inclined than promise-breakers and distrusted trustees to construct messages for the purpose of persuading investors to choose IN in Game 2, and so would send both shorter messages and a higher proportion of empty messages in Game 2. Supporting that expectation, Game 1 promise-keepers' messages contained fewer words than those from the set including both Game 1 distrusted trustees and Game 1 promise-breakers ($M_1 = 11.41$ ($SD_1 = 11.94$) versus $M_2 = 22.9$ ($SD_2 = 22.37$); Wilcoxon rank-sum test, $p < 0.01$, $n_1 = 155$, $n_2 = 74$). Game 1 promise-keepers' messages were also more frequently empty than those from the set of both Game 1 distrusted trustees and Game 1 promise-breakers (20.6% versus 10.8% of the time; Fisher's exact test, p -value = 0.05, $n = 229$).

Investments in Game 1 promise-keepers paid off for investors choosing IN in Game 2. These investors received an average of \$8.62 from trustees, as opposed to the \$5 earned from OUT (Wilcoxon signed rank test, p -value < 0.01, $n = 143$), with 83.9%

Fig. 3. Distribution of promises in Game 2 (resulting in IN or OUT) by Game 1 promise-keepers.

Table 3

Game 2 determinants of investment and return with Game 1 promise-keepers.

Dependent variable Specification	Invest2 (1) Probit	Invest2 (2) Probit	Return2 (3) OLS	Return2 (4) OLS
<i>Return1</i> [return in Game 1]	0.15 (0.15)	0.11 (0.15)	−0.01 (0.21)	−0.05 (0.21)
<i>Message</i> [message with content]	0.63* (0.38)	0.61 (0.38)	−0.24 (0.63)	−0.24 (0.63)
<i>Promise2 < 9</i> [dummy for promises less than 9 in Game2]	−1.12*** (0.43)	−4.39** (2.21)	−2.23*** (0.69)	−9.13 (5.57)
<i>Promise2 < 9 × Promise2</i> [amount promised in Game2 for promises < 9]		0.45 (0.30)		0.92 (0.74)
<i>Promise2 > 11</i> [dummy for promises greater than 11 in Game2]	−1.83** (0.80)	−1.82** (0.81)	9.12*** (2.01)	9.08*** (2.00)
<i>Constant</i>	0.29 (1.42)	0.63 (1.47)	9.23*** (2.04)	9.58*** (2.06)
Observations	155	155	143	143

Standard errors in parenthesis.

* Significant at 10%.

** Significant at 5%.

*** Significant at 1%.

(120/143) of the promises kept or exceeded, and 16.1% (23/143) broken. Compared to \$0 earned by those Game 1 promise-keepers that were not trusted in Game 2, promise-keepers also profited from trusted promises in Game 2, earning \$11.38 on average (Wilcoxon rank-sum test, p -value < 0.01, $n_1 = 12$, $n_2 = 143$). Specifications (3) and (4) in Table 3 indicate that Game 2 promises lower and higher than the even-split are associated with lower and higher amounts returned by Game 1 promise-keepers, respectively. Whether the messages have content or not, on the other hand, has no effect on returns.

4.2.2. Promise-breakers

A major question our data address concerns what happens after a breach of trust when a fresh opportunity for cooperation arises: How trustees behave, how investors respond, and what outcomes are achieved. Here we focus on the 18.8% (36/191) of pairs with broken promises in Game 1 (i.e. where the amount returned was lower than the amount promised). These broken promises represent breaches of trust and the relationships that immediately follow are considered to have *damaged trust* (because trust-based expectations were not met). A central question motivating this study is whether signals such as new promises and apologies can (i) restore investors' willingness to trust and (ii) facilitate the achievement of higher joint payoffs.

Fig. 4 displays the distribution of promises made by 36 Game 1 promise-breakers in Game 2 resulting in IN and OUT decisions. Promise-breakers promised \$12.11 in Game 2, which is significantly higher than their average promise of \$10.58 in Game 1 (Wilcoxon signed rank test, p -value = 0.01, $n_1 = n_2 = 36$). The extent of upgraded promises (*Promise2*–*Promise1*) by promise-breakers is also significantly higher than the extent of upgraded promises by promise-keepers ($M_1 = 1.53$ ($SD_1 = 3.70$) versus $M_2 = 0.44$ ($SD_2 = 1.79$); Wilcoxon rank-sum test, p -value < 0.01, $n_1 = 36$, $n_2 = 155$). Assuming that many of

Fig. 4. Distribution of promises in Game 2 (resulting in IN or OUT) by Game 1 promise-breakers.

Table 4

Game 2 determinants of investment and return with Game 1 promise-breakers.

Dependent variable Specification	Invest2 (1) Probit	Invest2 (2) Probit	Return2 (3) OLS	Return2 (4) OLS
<i>Return1</i>	0.10	−0.11	0.69***	1.15**
[return in Game 1]	(0.08)	(0.15)	(0.24)	(0.46)
<i>Promise2–Promise1</i>	0.21*		−0.46	
[promise adjustment]	(0.13)		(0.45)	
<i>Promise1–Return1</i>		−0.21*		0.46
[extent of under–return relative to Game1 promise]		(0.13)		(0.45)
<i>Apology</i>	1.88***	1.88***	2.69	2.69
[message with apology]	(0.68)	(0.68)	(3.19)	(3.19)
<i>Promise2</i>	−0.09	0.13	0.37	−0.09
[amount promised in Game 2]	(0.12)	(0.08)	(0.48)	(0.29)
<i>Constant</i>	−0.56	−0.56	−1.90	−1.90
	(1.58)	(1.58)	(6.24)	(6.24)
Observations	36	36	25	25

Standard errors in parenthesis.

* Significant at 10%.

** Significant at 5%.

*** Significant at 1%.

the investors whose trust had previously been damaged would be inclined to choose OUT, it appears that promise upgrades partially restore trust, since 69.4% (25/36) of investors whose trust was damaged in Game 1 chose IN again.

In addition to promise upgrades, we also find that Game 1 promise-breakers frequently used messages. Table C1 reports all messages that were sent by 36 promise-breakers. Analyzing the messages, we find that 83.3% (30/36) of the messages have some content.⁶ Game 1 promise-breakers' messages contain more words than messages from Game 1 promise-keepers ($M_1 = 19.06$ ($SD_1 = 19.03$) versus $M_2 = 11.41$ ($SD_2 = 11.94$); Wilcoxon rank-sum test, p -value = 0.03, $n_1 = 36$, $n_2 = 155$), suggesting that the behavioral cue of trustworthiness made verbal persuasion for continued investment less determinant of re-investment and less necessary for trustees.

To further classify all 36 messages, we used an incentivized laboratory coordination game (Houser and Xiao, 2011). Three coders recruited from the subject pool and blind to the hypotheses⁷ were asked to code each message twice: first based on whether or not it conformed to a "broad" definition of apology (an explicit or implicit acknowledgment of offense), and second based on whether or not it conformed to a "narrow" definition of apology (an explicit or implicit acknowledgment of offense, along with remorse, regret, or sorrow stemming from acknowledgment of the offense). All 6 messages without content were coded by all coders as not conforming to the broad definition and not conforming to the narrow definition of apology. Of the 30 messages with content, 28 were coded by the majority of coders as conforming to the broad definition of apology and 13 were coded by the majority of coders as conforming to the narrow definition of apology.⁸

When using a broad definition of apology, which was coded with "substantial" agreement (Kappa of 0.70), we find that 82.1% (23/28) of apologizers were retrusted in Game 2 in comparison to only 25.0% (2/8) of non-apologizers (Fisher's exact test, p -value < 0.01, $n = 36$), suggesting that messages with apology are more likely to restore trust after broken promises than empty messages or messages without apology.⁹ Due to the lower interrater reliability for messages coded according to the narrow definition, we will consider only the broad definition of apology in the subsequent analyses.

Thus far, we have only considered the independent effects of new promises and apologies in restoring damaged trust, but recognize that these remedial strategies are often used jointly. Among Game 1 promise-breakers, the size of the upgrade in amount promised is significantly larger for participants issuing apologies than for those who did not ($M_1 = \$1.68$ ($SD_1 = 3.10$) versus $M_2 = \$1$ ($SD_2 = 5.55$); Wilcoxon rank-sum test, p -value = 0.05, $n_1 = 28$, $n_2 = 10$). When the apologetic promise-breakers are compared to all other trustees the difference is even larger. The upgrade in amount promised for apologetic trustees is almost four times greater than among all other trustees ($M_1 = \$1.68$ ($SD_1 = 3.10$) versus $M_2 = \$0.44$ ($SD_2 = 2.68$); Wilcoxon rank-sum test, p -value = 0.05, $n_1 = 28$, $n_2 = 201$), indicating that apologetic trustees upgraded their promises most.

Next, we estimate probit regressions (see Table 4) to identify how these remedial strategies work in conjunction. Specification (1) indicates that the two most significant predictors of trust in Game 2 are promise adjustments (specifically promise upgrades) and apologies. Specification (2) shows that in addition trust is negatively affected by the extent of under-

⁶ We find that 80% (24/30) of messages with content restored trust (i.e. where investors chose IN in Game 2 after having suffered broken promises in Game 1), as opposed to only 16.7% (1/6) of messages without content. These differences are significant (Fisher's exact test, p -value < 0.01, $n = 36$).

⁷ The instructions for coders and details about how they were paid are attached in Appendix A. Coders each earned an average of \$28.33 for matched codings, plus \$7 for arriving on time and participating.

⁸ We use a standard approach from content analysis methodology to calculate Cohen's Kappa interrater agreement coefficient (Cohen, 1960; Krippendorff, 2004). Kappa values between 0.41 and 0.60 are considered "Moderate" agreement, and those above 0.60 indicate "Substantial" agreement (Landis and Koch, 1977). We find Kappa values of 0.70 and 0.53 for the broad and narrow definitions of apology, respectively.

⁹ When using a narrow definition of apology, which was coded with moderate agreement (Kappa of 0.53), we find that 84.6% (11/13) of apologizers were retrusted in comparison to only 60.9% (14/23) of non-apologizers (Fisher's exact test, p -value = 0.13, $n = 36$).

return relative to Game 1 promise. Moreover, based on the likelihood-ratio test, we find that the promise adjustment, upgraded relative to the extent of under-return on previously broken promise (i.e. $\text{Promise2} - (\text{Promise1} - \text{Return1})$), positively and significantly influences trust in Game 2 (likelihood-ratio test, p -value = 0.05). These results indicate that investors' decisions to re-invest are sensitive not just to the existence of broken promises (specifications 1 and 2 in Table 2), but also to the extent of under-return relative to Game 1 promise, apologies, and upgraded promises (specifications 1 and 2 in Table 4).

We have argued that signals such as apologies and promises should have evolved only if they provided net benefits to both the senders and receivers of those signals on average and over the evolution of the communication system. We evaluate whether Game 1 promise-breakers' signals resulted in benefits for both investor and trustee in Game 2, and whether these signals were reliable indicators of subsequent trustee behaviors. Investors in Game 1 promise-breakers were returned \$7.28 on average, which is significantly higher than the OUT payoff of \$5 (Wilcoxon signed rank test, p -value = 0.05, $n = 25$). Moreover, Game 1 promise-breakers returned significantly more in Game 2 than in Game 1 ($M_1 = \$7.28$ ($SD_1 = 4.86$) versus $M_2 = \$4.60$ ($SD_2 = 3.72$); Wilcoxon signed rank test, p -value < 0.01, $n_1 = n_2 = 25$). This is also true when we look at investments in the subset of 23 out of 28 trustees who issued apologies and where re-trusted: they returned significantly more in the second game ($M_1 = \$7.52$ ($SD_1 = 4.81$) versus $M_2 = \$4.61$ ($SD_2 = 3.64$); Wilcoxon signed rank test, p -value < 0.01, $n = 23$), which is also significantly higher than the OUT payoff ($M_1 = \$7.52$ ($SD_1 = 4.81$) versus $M_2 = \$5$; Wilcoxon signed rank test, p -value = 0.03, $n = 23$).

Although investments in Game 2 paid off, we still find that 60.0% (15/25) of re-trusted promise-breakers subsequently broke their promises again in Game 2 – almost irrespective of the apologies and new adjusted promises. From specifications (3) and (4) in Table 4, it appears that neither promises adjustments, new amounts promised, nor apologies are predictive of return in Game 2. The only variable that predicts return in Game 2 is return in Game 1.

4.2.3. Distrusted

As mentioned above, 16.6% of trustees (38 out of 229) were not trusted in Game 1 (see Fig. 1). The source of this distrust appears to be: (i) the higher variance around the even-split point of the distribution of distrusted promises (relative to that of the distribution of trusted promises; see right panel of Fig. 2), and (ii) a lower degree of default trustfulness among Game 1 distrustful investor (accounting for the fact that a sizeable number of even-split promises were rejected). In particular, in Game 1, 55.3% (21/38) of distrusted trustees promised less than \$9 while another 10.5% (4/38) of them promised more than \$11. As with our Game 1 predictions of trusted promises, we expect that previously distrusted trustees would adjust their Game 2 promises towards the modal and more trusted promise of \$10, that these adjustments would affect decisions to invest, and that investments made based on adjusted promises would benefit both the investor and trustee.

First we evaluate whether Game 1 distrusted trustees adjust their promises as we expected, and if adjustments of promises by Game 1 distrusted trustees affect investment decisions. Trustees who were distrusted in Game 1 promised an average of \$8.92 in Game 2, which is similar to their average promise of \$8.61 in Game 1 (Wilcoxon signed rank test, p -value = 0.45, $n_1 = n_2 = 38$), yet most investors (84.2% or 32/38) who did not trust in Game 1 chose IN in Game 2. Fig. 5 displays the histogram of promises made in Game 2 by the 38 trustees who were distrusted in Game 1. Distrusted trustees changed their distribution of promises towards more equal splits: 66.7% (14/21) of trustees who promised less than \$9 in Game 1 increased their Game 2 promises and 100% (4/4) of trustees who promised more than \$11 in Game 1 decreased their Game 2 promises. Correspondingly, among previously un-trusting investors, 92.6% (13/14) of those who received increased promises and 100% (4/4) of those who received the decreased promises chose IN in Game 2.

Fig. 5. Distribution of promises in Game 2 (resulting in IN or OUT) by Game 1 distrusted trustees.

Table 5

Game 2 determinants of investment and return with Game 1 distrusted trustees.

Dependent variable Specification	<i>Invest2</i> (1) Probit	<i>Return2</i> (2) OLS
<i>Promise1 = 10</i> [dummy for promises equal to 10 in Game1]	−0.86 (0.70)	2.51 (2.16)
<i>Wordcount</i> [number of words]	0.02 (0.01)	0.00 (0.04)
<i>Promise2 = 10</i> [dummy for promises equal to 10 in Game2]	1.40 [*] (0.76)	0.08 (1.89)
<i>Constant</i>	0.38 (0.47)	6.28 ^{***} (1.73)
Observations	38	32

Standard errors in parenthesis.

^{*} Significant at 10%.^{**} Significant at 5%.^{***} Significant at 1%.

Next, we analyze whether new trust in previously distrusted trustees can be statistically attributed to how distrusted trustees utilized messages and adjusted promises. We expect that distrusted trustees would construct longer messages (to persuade investors to choose IN in Game 2) than trustees who had already established reputations of trustworthiness. Table C2 reports the messages that were sent by 38 trustees who were distrusted in Game 1. Analyzing these messages, we find that 94.7% (36/38) of the messages used by distrusted trustees have some content. Messages from Game 1 distrusted trustees contain more words than messages from Game 1 promise-keepers ($M_1 = 26.58$ ($SD_1 = 24.83$) versus $M_2 = 11.41$ ($SD_2 = 11.94$); Wilcoxon rank-sum test, $p < 0.01$, $n_1 = 38$, $n_2 = 155$). These data suggest that distrusted trustees use both promise adjustments towards 50/50 divisions of income and longer messages to persuade investors to trust them. The estimation of specification (1) in Table 5 indicates that the investment decision in Game 2 is positively correlated with the amount of \$10 promised in Game 2 (p -value = 0.06), but that message length is not significant. The Game 2 rate of trust-extension for Game 1 distrustful investors was 84.2% (32/38). This is very similar to the original unconditional investment rate of 83.4% in Game 1.

Finally, we evaluate whether promises used by Game 1 distrusted trustees facilitated higher joint payoff in Game 2, and whether those promises were reliable indications of subsequent trustee behavior. Game 2 investments made in previously distrusted trustees paid off for both investors and trustees. Investors in Game 1 distrusted trustees were returned on average \$6.88, which is significantly higher than the OUT payoff of \$5 (Wilcoxon signed rank test, p -value = 0.05, $n = 32$) and the newly trusted trustees earned an average of \$13.12.

On average, newly trusted trustees' promises were veridical – 62.5% (20/32) kept their promises. On the other hand, 37.5% (12/32) of that set broke their promises – more than the 18.8% of trusted trustees who broke their promises in Game 1 (Wilcoxon rank-sum test, $p = 0.05$, $n_1 = 32$, $n_2 = 191$). One possibility is that the excess promise-breaking of newly trusted trustees reflects a reaction against investors' lack of trust in Game 1, perhaps based on a sense of entitlement for the profits that could have been earned had trust been extended in Game 1. By breaking their promises these presumed punishers ended up earning an average of \$17.42 in two games, closer to the average earning of \$21.99 across two games for Game 1 trusted trustees, than the average earnings of \$10.55 for newly trusted trustees who did not break promises in Game 2. The estimation of specification (2) in Table 5 does not reveal any significant predictor for amount returned by trustees in Game 2.

5. Discussions and conclusions

Opportunities for mutual gains often exist where previous exchange histories have not yet been developed or where trust has been damaged by expectations not met. While promises and apologies appear to be important tools for building and rebuilding trust in these problematic situations, most of the research on these remedial strategies is based on self-report, anecdotal, or archival evidence, or else experiments based on fictional vignettes, videotaped dramatization, or deception. By using a non-deceptive study wherein financially motivated participants used endogenously created and naturally distributed promises and apologies we demonstrate that trustees send cheap signals to encourage new trust and rebuild damaged trust, and that these signals are often effective, leading to benefits for both investor and trustee.

From the egoist perspective of non-cooperative game theory no cooperation is predicted, yet our experiment yielded high rates of trust-extending behavior (e.g. 83.4% in Game 1) and trust-reextension (88% of those who went IN in Game 1 went IN in Game 2). There are several non-exclusive accounts for these results. Profit-seeking investors need to trade the risk of trusting under-reciprocators against the risk of not trusting reciprocators.¹⁰ As the efficiency of the investment increases,

¹⁰ In Game 1, for instance, investors who chose IN received back \$8.19 on average, which is significantly higher than their alternative OUT payoff of \$5. Despite the sizeable rate of under-reciprocation in Game 1, decisions to take on risk by choosing IN yielded higher profits for investors than decisions to not take on risk by choosing OUT.

so do the possible forgone benefits for investors who choose OUT. While the multiplier of 4 used in our study, higher than the multiplier of 3 used in standard trust games, might have contributed to investor willingness to choose IN, these effects are not commonly found across trust games. A meta-analysis by Johnson and Mislin (2011), examining games with different multipliers to evaluate whether a higher multiplier might increase the likelihood of investment, found no effect of multiplier on investors and a strong negative effect on trustworthiness: a higher multiplier *decreases* the amount of money returned by a receiver.

Another possibility, which our data suggests, is that promises in our game may have enhanced the trust extension rate (consistent with Charness and Dufwenberg, 2006, who showed IN rates of 74% with promises), and that the apologies issued may have enhanced the trust re-extension rate, despite broken promises in Game 1. Although we lack experimental controls without signals, within sample comparisons suggest that investors lent credence to specific formulations of trustees' promises, and conditioned their credence to these promises in Game 2 according to cues of trustworthiness and untrustworthiness (i.e. kept promises and broken promise, respectively). We expected that trustees, aware of investor self-interest and motives for critical signal reception, would promise investors mutually beneficial transfers in the range \$6–\$19. Consistent with results from bargaining games (where even splits are reported as modal offers that also tend to be accepted; see Guth et al., 1982; Guth and Tietz, 1986; Carter and Irons, 1991; Prasnikar and Roth, 1992), even-split promises, which lay close to the middle of the predicted range, elicited more trust-extension than uneven-split promises among dyads with no history of trust-based exchange.

We have argued that as long as the truth value of a signal can be reliably estimated, and updated in tandem with estimates of the signaler's trustworthiness, cheap-to-produce signals such as promises can facilitate coordination and cooperation. In the context of repeated interactions, promises and apologies should be less trusted when issued by trustees whose past promises and apologies were followed by untrustworthy behavior. The proverbial boy who cried wolf illustrates this principle in the domain of predator calls. But does the principle apply in the domain of social exchange? Our experimental design did not include a third game, so we cannot know whether the investors who suffered broken promises in Game 1, were apologized to, and again suffered broken promises in Game 2, would have discounted further apologies. Future research with similar designs but more than two successive games is needed to test the prediction that the credibility attributed to apologies would be recalibrated based on subsequent behavior by offenders. However, our results do provide a partial answer to the question of how signal credibility is calibrated by relevant behavioral cues: as evidenced by Game 2 investment rates, Game 2 promises issued by trustees who previously returned less than promised were given less credence than the Game 2 promises issued by trustworthy trustees.

Nevertheless, among investors whose trust was damaged in Game 1, messages with apologies elicited more re-extension of trust than messages without apologies. While Game 2 promise upgrades might signal intention to provide an economic contribution towards restituting the previously promised amount lost (i.e. *atonement*: a repair done for the sake of a damaged relationship), they could also be attempts at *coercion*: promise-breakers' coaxing efforts – calibrated to compensate for the investor's expectation that their Game 2 promises would be “exaggerated” (i.e. as they were in Game 1). The former explanation suggests an upgraded regard for the investor, the latter a strategically selfish regard. Evidence from our experiment suggests that most promise breakers upgraded their Game 2 promises out of selfish-regard, since the majority (60%) of promise-breakers who were invested in again went on to break promises a second time.

We suggest that the rate of trust re-extension seen for trustees who turned out to be repeat promise-breakers may have been lower outside of the laboratory, where emotional states are reliably communicated through other forms simultaneously (e.g. facial expressions, voice, body language) and in concert with additional reputational information and opportunities to sanction cheating. We suspect that in the “real world” of non-anonymous and face-to-face interactions, persuasive messages like promises and apologies are likely more effective and less likely to lead to further damaged trust because receivers can evaluate the veracity of verbal messages based on their correspondence with other reliable cues and signals (e.g. past demonstrations of trust or trustworthiness, facial expressions, tone of voice, eye movements, body language).¹¹

From this study we see evidence indicating how personal exchanges are often based around establishment of trust via cheap-to-produce verbal signals, and how these signals can encourage new trust where it did not previously exist or repair trust where it had been damaged. Not only is this important information that could improve understanding of what to expect from our everyday interpersonal relationships, it is information that complements our understanding of how market exchange systems (where interactions often take place between non-personal entities such as firms), politics, law, and religion are sometimes expected to work, with personal representatives making verbal and written promises of reciprocation or atonement or else issuing apologies and personalized messages. Both interpersonal interactions and markets are built on the ancient human foundations of adaptive giving and receiving. As such, trust-based exchanges at any level are often based around establishment of trust via signals such as verbal claims about reputation, verbal contracts, and apologies.

¹¹ Hirschleifer (1984) theorized that emotions act as “guarantors of threats and promises” and several authors (Van Kleef et al., 2004, 2006; Sinaceur and Tiedens, 2006; Wubben et al., 2009; Stouten and De Cremer, 2010) have demonstrated experimentally that displays of emotion (including anger, guilt, happiness, disappointment, worry, regret) are used by observers for subsequent decision making in social dilemmas and negotiations.

Based on our findings and a review of the current literature we suggest three steps that can be taken as a remedial strategy to restore damaged trust. First, when trust in a relationship has been damaged, the offender should recognize the offense and any regret or sorrow stemming from having caused the offense (such as through some form of apology). An optimistic perspective on relationships fraught with damaged trust recognizes that they actually represent opportunities to develop better relations than previously established. Second, to persuade and assure victims that relationship repair is possible the offender should signal (such as with a personalized message) an indication of the value that is recognized in the other, which stems from an internal recalibrations and commits one (such as with a promise) to expectations of future cooperative behavior. In signaling recognition of relationship value it is important not to express a selfish perspective, but instead a shared-welfare or other-regarding perspective. Third, to actually begin the process of changing and redefining the relationship, an offender must be willing to pay costs to expeditiously correct the previous imbalance of welfare (thereby increasing the offended party's welfare), or else to sacrifice wealth or status (thereby decreasing one's own welfare). When corrective actions cannot be immediately taken, signals of intent to take corrective actions should be used. These three steps are identified as each having independent effects of improving impressions of the offender (Scher and Darley, 1997; Schlenker, 1980) and are consistent with the proscriptions detailed by De Cremer (2010) for the financial world to restore their damaged trust with customers, as well as the general conclusions arrived at by Lazare (2004).

As the natural occurrence of deceit in social exchanges is sampled and the effectiveness of strategies, tools, and institutions used to combat it are evaluated, practical insights are gleaned that can be extended to our personal lives, to the work of policy makers, and even applied to the practices of firms, religious clergy, and military relations. We strongly encourage further efforts to uncover effective strategies for building trust where previous trust-based exchange histories had not been developed, or where trust had been damaged by reciprocity failure.

Appendices A–C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jebo.2012.09.011>.

References

- Axelrod, R., Dion, D., 1988. The further evolution of cooperation. *Science* 242, 1385–1389.
- Balliet, D., 2010. Communication and cooperation in social dilemmas: a meta-analytic review. *Journal of Conflict Resolution* 54 (1), 39–57.
- Barnett, M., 2003. Unringing the bell: can industries reverse unfavorable institutional shifts triggered by their own mistakes? In: *Southern Management Association Conference Proceedings*, pp. 800–806.
- Benoit, W.L., Drew, S., 1997. Appropriateness and effectiveness of image repair strategies. *Communication Reports* 10, 153–163.
- Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity and social history. *Games and Economic Behavior* 10, 122–142.
- Bohnet, I., Frey, B.S., 1999. The sound of silence in prisoner's dilemma and dictator games. *Journal of Economic Behavior and Organization* 38, 43–57.
- Bottom, W., Daniels, S., Gibson, K.S., Murnighan, J.K., 2002. When talk is not cheap: substantive penance and expressions of intent in the reestablishment of cooperation. *Organization Science* 13, 497–515.
- Bracht, J., Feltovich, N., 2009. Whatever you say, your reputation precedes you: observation and cheap talk in the trust game. *Journal of Public Economics* 93, 1036–1044.
- Buchan, N.R., Croson, R., Johnson, E.J., 2006. Let's get personal: an international examination of the influence of communication, culture, and social distance on other regarding preferences. *Journal of Economic Behavior and Organization* 60 (3), 373–398.
- Carter, J.R., Irons, M.D., 1991. Are economists different, and if so, why? *Journal of Economic Perspectives* 5, 171–177.
- Charness, G., Dufwenberg, M., 2006. Promises and Partnership. *Econometrica* 74, 1579–1601.
- Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37–46.
- Conway, N., Briner, R.B., 2002. A daily diary study of affective responses to psychological contract breach and exceeded promises. *Journal of Organizational Behavior* 23, 287–302.
- Dawkins, R., Krebs, J.R., 1978. Animal signals: information or manipulation. In: Krebs, J.R., Davies, N.B. (Eds.), *Behavioural Ecology: An Evolutionary Approach*, 1st ed. Blackwell Scientific, Oxford, pp. 282–309.
- De Cremer, D., 2010. Rebuilding trust. *Business Strategy Review* 21 (2), 79–80.
- De Cremer, D., van Dijk, E., Pillutla, M.M., 2010. Explaining unfair offers in ultimatum games and their effects on trust: an experimental approach. *Business Ethics Quarterly* 20 (1), 107–126.
- Delton, A.W., Krasnow, M.M., Cosmides, L., Tooby, J., 2011. Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences of the United States of America* 108, 13335–13340.
- Deutsch, M., 1958. Trust and suspicion. *Journal of Conflict Resolution* 2, 256–279.
- Deutsch, M., 1960. The effect of motivational orientation upon trust and suspicion. *Human Relations* 13, 123–133.
- Diggle, S.P., Gardner, A., West, S.A., Griffin, A.S., 2007. Evolutionary theory of bacterial quorum sensing: when is a signal not a signal? *Philosophical Transactions of the Royal Society B* 362, 1241–1249.
- Dirks, K.T., Lewicki, R.J., Zaheer, A., 2009. Repairing relationships within and between organizations: building a conceptual foundation. *Academy of Management Review* 34 (1), 68–84.
- Duffy, J., Feltovich, N., 2006. Words, deeds, and lies: strategic behavior in games with multiple signals. *Review of Economic Studies* 73, 669–688.
- Elingsen, T., Johannesson, M., 2004. Promises, threats, and fairness. *Economic Journal* 114, 397–420.
- Elsbach, K.D., 1994. Managing organizational legitimacy in the California cattle industry: the construction and effectiveness of verbal accounts. *Administrative Science Quarterly* 39, 57–88.
- Farrell, J., Rabin, M., 1996. Cheap talk. *Journal of Economic Perspectives* 10 (3), 103–118.
- Fischbacher, U., Utikal, V., 2010. On the Acceptance of Apologies. Thurgau Institute of Economics and Department of Economics, Research Paper Series No. 53.
- Gibson, K., Bottom, W., Murnighan, K., 1999. Once bitten: defection and reconciliation in a cooperative enterprise. *Business Ethics Quarterly* 9, 69–85.
- Gillespie, N., Dietz, G., 2009. Trust repair after an organization-level failure. *Academy of Management Review* 34, 127–145.
- Girard, M., Mullet, E., 1997. Propensity to forgive in adolescents, young adults, older adults, and elderly people. *Journal of Adult Development* 4, 209–220.
- Girard, M., Mullet, E., Callahan, S., 2002. The mathematics of forgiveness. *American Journal of Psychology* 115, 351–375.
- Grafen, A., 1990. Biological signals as handicaps. *Journal of Theoretical Biology* 144, 517–546.

- Guth, W.R., Schmittberger, W.R., Schwartz, B., 1982. An experimental analysis of ultimatum games. *Journal of Economic Behavior and Organization* 3, 367–388.
- Guth, W.R., Tietz, P., 1986. Acutoning ultimatum bargaining positions: how to act if rational decisions are unacceptable. In: Scholz, R.W. (Ed.), *Current Issues in West German Decision Research*. P. Lang, Frankfurt, pp. 173–185.
- Haselton, M.G., Nettle, D., 2006. The paranoid optimist: an integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review* 10 (1), 47–66.
- Hirshleifer, J., 1984. On the emotions as guarantors of threats and promises. UCLA Department of Economic Working Paper #337, August.
- Hoffman, E., McCabe, K., Smith, V.L., 1996. Social distance and other-regarding behavior in dictator games. *American Economic Review* 86 (3), 653–660.
- Houser, D., Xiao, E., 2011. Classification of natural language messages using a coordination game. *Experimental Economics* 14, 1–14.
- Johnson, N.D., Mislin, A.A., 2011. Trust games: a meta-analysis. *Journal of Economic Psychology* 32, 865–889.
- Kelly, R.L., 1995. *The Foraging Spectrum*. Smithsonian Institution Press, Washington, DC.
- Kerr, N., Kaufman-Gilliland, C., 1994. Communication, commitment, and cooperation in social dilemmas. *Journal of Personality and Social Psychology* 66, 513–529.
- Kim, P.H., Ferrin, D.L., Cooper, C.D., Dirks, K.T., 2004. Removing the shadow of suspicion: the effects of apology vs. denial for repairing ability- vs. integrity-based trust violations. *Journal of Applied Psychology* 89, 104–118.
- Kim, P.H., Dirks, K.T., Cooper, C.D., 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, 49–65.
- Kiyonari, T., Tanida, S., Yamagishi, T., 2000. Social exchange and reciprocity: confusion or a heuristic? *Evolution and human behavior* 21 (6), 411–427.
- Knight, R.F., Pretty, D.J., 1999. Corporate catastrophes, stock returns, and trading volume. *Corporate Reputation Review* 2 (4), 363–378.
- Krippendorff, K., 2004. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Thousand Oaks, CA.
- Lachmann, M., Számadó, S., Bergstrom, C.T., 2001. Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences* 98 (23), 13189–13194.
- Landis, J.R., Koch, G., 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 33, 363–374.
- Lazare, A., 2004. *On Apology*. Oxford University Press, New York.
- Lewicki, R.J., Wiethoff, C., 2000. Trust, trust development, and trust repair. In: Deutsch, M., Coleman, P.T. (Eds.), *Handbook of Conflict Resolution: Theory and Practice*. Jossey-Bass, San Francisco, CA.
- Lewicki, R.J., Bunker, B.B., 1996. Developing and maintaining trust in working relationships. In: Kramer, R.M., Tyler, T.R. (Eds.), *Trust in Organizations: Frontiers of Theory and Research*. Sage, Thousand Oaks, CA.
- Masclot, D., Noussair, C., Tucker, S., Villeval, M., 2003. Monetary and non-monetary punishment in the voluntary contributions mechanism. *The American Economic Review* 93 (1), 366–380.
- Maynard-Smith, J., 1982. Do animals convey information about their intentions? *Journal of Theoretical Biology* 97, 1–5.
- McCullough, M.E., Rachal, K.C., Sandage, S.J., Worthington, E.L., Brown, S.W., Hight, T.L., 1998. Interpersonal forgiving in close relationships. II. Theoretical elaboration and measurement. *Journal of Personality and Social Psychology* 75, 1586–1603.
- McCullough, M.E., Worthington Jr., E.L., Rachal, K.C., 1997. Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology* 73, 321–336.
- Mishra, A.K., 1996. Organizational responses to crisis: the centrality of trust. In: Kramer, R.M., Tyler, T.R. (Eds.), *Trust in Organizations: Frontiers of Theory and Research*. Sage, Thousand Oaks, CA, pp. 261–287.
- Nakayachi, K., Watabe, M., 2005. Restoring trustworthiness after adverse events: the signaling effects of voluntary “hostage posting” on trust. *Organizational Behavior and Human Decision Processes* 97, 1–17.
- Ohbuchi, K., Kameda, M., Agarie, N., 1989. Apology as aggression control: its role in mediating appraisal of and response to harm. *Journal of Personality and Social Psychology* 56, 219–227.
- Ohtsubo, Y., Watanabe, E., 2009. Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior* 30 (2), 114–123.
- Orbell, J., Dawes, R., Kragt van de, A., 1988. Explaining discussion induced cooperation. *Journal of Personality and Social Psychology* 54, 811–819.
- Ostrom, E., Walker, J., 2005. *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*. Russell Sage Foundation Publications.
- Petrie, M., Halliday, T., Sanders, C., 1991. Peahens prefer peacocks with elaborate trains. *Animal Behavior* 41 (2), 323–331.
- Prasnikar, V., Roth, E., 1992. Considerations of fairness and strategy: experimental data from sequential games. *Quarterly Journal of Economics* 106, 865–888.
- Radlow, R., Weidner, M.F., 1966. Unenforced commitments in “cooperative” and “noncooperative” non-constant-sum games. *Journal of Conflict Resolution* 10, 497–550.
- Ren, H., Gray, B., 2009. Repairing relationship conflict: how violation types and culture influence the effectiveness of restoration rituals. *Academy of Management Review* 34, 105–126.
- Ridings, C.M., Gefen, D., Arinze, B., 2002. Some antecedents and effects of trust in virtual communities. *Journal of Strategic Information Systems* 11 (3), 271–295.
- Robinson, S.L., Rousseau, D.M., 1994. Violating the psychological contract: not the exception but the norm. *Journal of Organizational Behavior* 15, 245–259.
- Rohwer, S., 1977. Status signaling in Harris sparrows: some experiments in deception. *Behaviour* 61 (1/2), 107–129.
- Rousseau, D.M., 1989. Psychological and implied contracts in organizations. *Employee Responsibilities and Rights Journal* 2 (2), 121–139.
- Rubin, J., Brown, B., 1975. *The Social Psychology of Bargaining and Negotiation*. Academic Press, New York, NY.
- Sally, D., 1995. Conversation and cooperation in social dilemmas. *Rationality and Society* 7 (1), 58–92.
- Scher, S.J., Darley, J.M., 1997. How effective are the things people say to apologize? Effects of the realization of the apology speech act. *Journal of Psycholinguistic Research* 26, 127–140.
- Schlenker, B.R., 1980. *Impression Management: The Self-concept, Social Identity, and Interpersonal Relations*. Brooks/Cole Publishing Company, Monterey, CA.
- Schweitzer, M.E., Hershey, J., Bradlow, E., 2006. Promises and lies: restoring violated trust. *Organizational Behavior and Human Decision Processes* 101, 1–19.
- Scott-Phillips, T.C., 2008. Defining biological communication. *Journal of Evolutionary Biology* 21, 387–395.
- Sinaceur, M., Tiedens, L.Z., 2006. Get mad and get more than even: when and why anger expression is effective in negotiations. *Journal of Experimental Social Psychology* 42, 314–322.
- Slovic, P., 1993. Perceived risk, trust, and democracy. *Risk Analysis* 13, 675–682.
- Stouten, J., De Cremer, D., 2010. Seeing is believing: the effects of facial expressions of emotion and verbal communication in social dilemmas. *Journal of Behavioral Decision Making* 23, 271–287.
- Tavuchis, N., 1991. *Mea Culpa: A Sociology of Apology and Reconciliation*. Stanford University Press, Stanford, CA.
- Tomlinson, E.C., Dineen, B.R., Lewicki, R.J., 2004. The road to reconciliation: antecedents of victim willingness to reconcile following a broken promise. *Journal of Management* 30, 165–187.
- Van Kleef, G.A., De Dreu, C.K.W., Manstead, A.S.R., 2004. The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology* 86, 57–76.
- Van Kleef, G.A., De Dreu, C.K.W., Manstead, A.S.R., 2006. Supplication and appeasement in conflict and negotiation: the interpersonal effects of disappointment, worry, guilt and regret. *Journal of Personality and Social Psychology* 91, 124–142.

- Witvliet, C.V.O., Ludwig, T.E., Bauer, D.J., 2002. Please forgive me: transgressors' emotions and physiology during imagery of seeking forgiveness and victim responses. *Journal of Psychology and Christianity* 21, 219–233.
- Wubben, M.J.J., De Cremer, D., van Dijk, E., 2009. When and how communicated guilt affects contributions in public good dilemmas. *Journal of Experimental Social Psychology* 45, 15–23.
- Zahavi, A., 1975. Mate selection: a selection for a handicap. *Journal of Theoretical Biology* 53, 205–214.
- Zahavi, A., 1993. The fallacy of conventional signalling. *Philosophical Transactions of the Royal Society London, Series B* 340, 227–230.
- Zheng, J., Veinott, E., Bos, N., Olson, J.S., Olson, G.M., 2002. Trust without touch: jumpstarting long-distance trust with initial social activities. In: *Proceedings of CHI*. ACM Press, New York, NY, pp. 141–146.