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

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),

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.