DOI:10.1093/esr/jcs078, available online at www.esr.oxfordjournals.org
Online publication 24 October 2012

# Temporal Embeddedness and Signals of Trustworthiness: Experimental Tests of a Game Theoretic Model in the United Kingdom, Russia, and Switzerland

Wojtek Przepiorka[1],* and Andreas Diekmann[2]

***Abstract:*** Based on signalling theory, Posner (1998, 2000) suggests that seemingly irrational behaviour or social norms emerge because they help to distinguish agents who prefer to engage in repeated cooperative interactions (long-term types) from agents with immediate non-cooperative incentives (short-term types). In this article, we formalize Posner's theory in a signalling trust game, derive hypotheses from our model, and test them empirically in a series of laboratory experiments in the United Kingdom, Russia, and Switzerland. Our results are surprisingly robust across the three countries and in line with most of our hypotheses. However, contrary to our main hypothesis, the introduction of a signalling opportunity does not increase the overall level of trust in our experiments and even reduces it under certain conditions. We argue that this is due to a high proportion of short-term types honouring trust because of non-selfish motives. Our results show that if *a priori* levels of trust and trustworthiness are high, introducing a signalling opportunity that is meant to distinguish long-term and short-term types may have a counterproductive effect.

## Introduction

Trust is an indispensable ingredient in cooperative social interactions. In everyday life, people trust each other even though such trust could be abused without any consequences for the person trusted (the trustee) and even though the person trusting (the truster) could incur considerable losses. Buying a used car, employing someone, falling in love, or paying a bribe are all examples of actions taken in the hope of mutual benefit but with the risk that the other party might have sufficient incentive to abuse the trust. The car dealer could sell you a lemon, an employee could quit after receiving valuable on-the-job training, a lover could steal away after one night, and the person who has taken the bribe could keep the money and provide no reciprocal service. If the truster fears that the trustee is going to abuse the trust, he or she will not trust in the first place.

### Trust and Uncertainty

Trust problems arise in sequential exchanges when the first mover, the truster, is uncertain about whether his or her utility transfer will be reciprocated by the second mover, the trustee. The exchange is not based on a formally binding agreement, and so a self-interested trustee has an incentive not to reciprocate. If the stakes are too high, the truster prefers not to place trust and both actors are worse off than if trust was given and honoured (Coleman, 1990).

The truster's uncertainty about whether the trustee is trustworthy follows from the assumption that trustees differ in the preferences and constraints that shape their incentives in the interaction (Riegelsberger, Sasse, and McCarthy (2005) distinguish between intrinsic and contextual trust-warranting properties). First, a trustee could have restrictive preferences. He or she might have the opportunity, but not the desire, to abuse trust.

[1]Department of Sociology, University of Oxford, Oxford OX1 3UQ, UK and [2]Chair of Sociology, ETH Zurich, Zurich CH-8092, Switzerland. *Corresponding author. Email: wojtek.przepiorka@sociology.ox.ac.uk

Submitted: September 2011; revised: July 2012; accepted: July 2012.

A trustee might prefer to be honest because of internalized norms of reciprocity and fairness (Voss, 1998; Bacharach and Gambetta, 2001) or because he or she would feel guilty for abusing the truster's trust (Braun, 1992; Snijders, 1996). Second, the trustee could be constrained by the structure of the interaction. In other words, either he or she might not have the option of abusing trust (Raub, 2004) or the payoffs might not provide sufficient incentive (Camerer and Weigelt, 1988). Third, a trustee could be concerned about the consequences in terms of his or her reputation in regard to the decision to abuse trust. Such social constraints result from the trustee's social embeddedness (Hardin, 1996; Cook and Hardin, 2001; Buskens and Raub, 2002).

A related argument suggested by Posner (1998, 2000) is that trustees differ in terms of their discount factors, *i.e.*, their probability of engaging repeatedly in cooperative interactions. In the simplest case, long-term types have high discount factors, and therefore a large 'shadow of the future' (Axelrod, 1984), and short-term types have low discount factors and thus immediate non-cooperative incentives. As discount factors are unobservable traits, there is uncertainty on the part of the truster about a potential trustee's type.

## Signalling Trustworthiness

A truster benefits from knowing who to trust and to distrust, while a trustee benefits from being trusted (irrespective of his or her type). Thus, honest and dishonest trustees have an incentive to offer credible information about their trustworthiness (*i.e.*, about their trustworthy-making preferences and constraints) or to convey a false impression of this, respectively. The question is how such information allows a truster to discriminate between the trustworthy type and the mimic.

Signalling theory (Spence, 1973; Voss, 1998; Bacharach and Gambetta, 2001; Raub, 2004) provides an answer. 'Signals' can be thought of as agents' observable actions conveying relevant information in a social interaction. The main tenets of signalling theory are as follows (see also Bliege Bird and Smith, 2005). First, individuals differ in the preferences and constraints determining their decisions in social interactions. Second, these preferences and constraints are only imperfectly observable, but are correlated with signalling costs and/or benefits such that certain individuals cannot afford to send signals at all, or only to a lesser extent (Johnstone, 1997; Gambetta, 2009). Third, individuals benefit from adequate information about their interaction partner's preferences and constraints. Finally, signals allow individuals to make inferences about these preferences and constraints, thereby reducing uncertainty.

For instance, a used car dealer selling lemons (unknown to the buyer) cannot afford to give a three-year guarantee to the buyer, a fresh employee already looking out for another job (unknown to the employer) will not bother to move near the employer's place of work, and a producer of low-quality goods (unknown to the consumer) will not spend money on advertising as dissatisfied consumers will not consider a second purchase (Nelson, 1974).

The last two examples are paradigmatic for Posner's idea that costly signals are indicative of agents' underlying discount factors. Both the committed employee as well as the high-quality producer are long-term types and thus have high discount factors. The former signals his or her 'long-term-ness' by incurring the potential costs of a removal, the latter by investing in his or her reputation through costly advertising. But Posner's theory goes a step further. It suggests that social norms are the equilibrium outcomes of such signalling games and adherence to social norms is a credible signal of one's long-term interest in cooperative relations. We expand on this idea in more detail elsewhere (Diekmann and Przepiorka, 2010). In this article, we focus on Posner's theory as suggesting that agents' temporal embeddedness is a trustworthy-making property that can be credibly signalled in social interactions when trust is at stake. The main question we want to answer here is whether long-term types invest in costly signals more and therefore are trusted more than short-term types and whether introducing a signalling opportunity increases the overall level of trust and collective gains.

## Experimental Evidence

To our knowledge, the only laboratory experiment with trust games and costly signals as a commitment device was conducted by Bolle and Kaehler (2007) (for other experiments on trust and commitments see, for example, Snijders and Buskens, 2001; Charness and Dufwenberg, 2006; Vieth, 2009). In their experiment, an 'honest' trustee gains from honouring a truster's trust, whereas a 'dishonest' trustee gains from abusing it. Subjects in the role of a truster cannot tell who is who, but they know that they will meet an honest trustee with probability $\alpha$. Given $\alpha$ and the possible payoffs from an interaction with either type, a rational and self-interested truster will decide whether to place trust based on the expected payoffs from either action. To be more precise, a theoretical threshold value $\alpha^*$ exists at which the expected payoffs from either of the truster's actions are equal and the truster is indifferent with regard to placing and not placing trust. In their experiment, Bolle and Kaehler (2007) vary $\alpha$. In one condition (low-alpha), $\alpha$ is below $\alpha^*$ and the truster has a higher expected payoff from not placing trust. In the other condition (high-alpha), $\alpha$ is above $\alpha^*$ and the truster has a higher expected payoff from placing trust.

In both conditions, the trustees are given the opportunity to signal their trustworthiness by making a costly gift to the truster before the truster decides whether to place trust. If the trustee decides to make a gift, he or she incurs the costs irrespective of what the truster does thereafter. The costs of the gift are such that even if the truster placed trust after receiving it, a dishonest trustee would not gain from having the opportunity to abuse that trust, whereas an honest trustee would still gain from honouring it. In other words, the signal is type-separating because only the honest trustee can afford to send it.

Such a signalling opportunity creates the conditions for a so-called 'separating equilibrium' in which the honest trustee sends a gift, a dishonest trustee does not (*i.e.*, type-separating behaviour), and the truster places trust only if he or she has received a gift. Obviously, the truster would benefit from knowing who to trust and who to distrust in both the low-alpha and the high-alpha condition. However, a separating equilibrium is more likely to emerge in the low-alpha condition than in the high-alpha condition because in the former, it is beneficial for the honest trustee to send a gift to be trusted, whereas in the latter, the truster is expected to place trust anyway and, therefore, the honest trustee can save the costs of a gift. In the high-alpha condition, it is therefore more likely that a so-called 'pooling equilibrium' emerges in which no trustee sends a gift (*i.e.*, type-pooling behaviour). Although their results did not perfectly align with these theoretical predictions, the results were relatively close.

Unlike in Bolle and Kaehler's experiment, the actors in our model are not distinguished by their payoffs in the trust game but by their discount factors, *i.e.*, by the expected number of trust games they can engage in. We outline our model in the next section.

# The Game Theoretic Model and Hypotheses

The signalling model described in this section involves a binary trust game (Dasgupta, 1988; Kreps, 1990), with the trustee holding private information about his or her type (*i.e.*, a trust game with incomplete information: see Camerer and Weigelt, 1988; Dasgupta, 1988; Voss, 1998; Buskens, 1999; Bacharach and Gambetta, 2001; Raub, 2004). In previous signalling models, which are based on the trust game with incomplete information, trustees' types differ either in the possibility or the immediate monetary incentives to abuse a truster's trust (Bacharach and Gambetta, 2001; Raub, 2004; Ahn and Esarey, 2008). Posner (1998, 2000) was the first to suggest that trustees can be distinguished based on their discount factors and he outlines the conditions for a separating equilibrium in a numerical example (Posner, 1998). However Posner's theory has been criticized on the grounds that it is difficult to prove empirically because potentially any time-, money-, or energy-consuming behaviour could serve as a signal (McAdams, 2001). To our knowledge, the model devised in this piece and the extended analysis contained in the online Supplementary Data 1 is the first attempt to formalize Posner's idea in game theoretic terms. We think that a formalized theory is more precise and allows hypotheses to be derived that can be tested empirically. Note, moreover, that our model is not restricted to cases in which trustees differ in terms of their discount factors; it can be equally applied to all cases in which one type of trustee prefers or is restricted to abuse the trust placed by a truster and another type of trustee prefers or is restricted to honour that trust. In the section 'Trust and Uncertainty', we listed several other reasons why trustees can differ in the preferences and constraints that shape their incentives to honour or abuse a truster's trust.

## Binary Trust Game

The binary trust game (Figure 1) represents a social dilemma that cannot be overcome by rational players. While it is rational for the truster not to place trust, both the truster and the trustee would be better off if trust was placed and honoured.

However, the assumption that trustees always abuse trust is neither realistic nor useful. If trusters were certain about the trustworthiness of trustees, the notion of trust would be superfluous. The trust problem arises from a truster's uncertainty about a trustee's preferences and constraints, which determine their decisions in an interaction. In the model presented here, we assume two types of trustees who differ in their discount factors and can be characterized as either long term or short term (Posner, 1998, 2000). A discount factor stands for a trustee's probability of engaging in another interaction with the truster, but it can also be interpreted as the trustee's time preference.[1] Given a long-term trustee's discount factor ($\delta_l$), he or she strictly prefers to engage in repeated interactions with a truster over a one-time abuse of trust [*i.e.*, $R/(1-\delta_l) > T > P$].[2] On the other hand, given a short-term trustee's discount factor ($\delta_s$), the expected payoff from repeated interactions with a truster is strictly smaller than his or her payoff from a

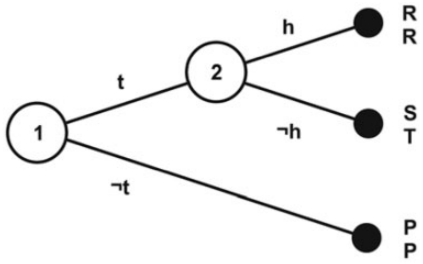

**Figure 1** In the binary trust game, the truster (player 1) first decides whether to place trust (t) or not to place trust (¬t). If the truster decides not to place trust, the interaction is terminated and both parties receive payoff P. If, instead, the truster decides to trust, it is the trustee's turn (player 2) to choose whether or not to honour that trust (h or ¬h). If the trustee honours the trust, both players receive payoff R. If the trustee does not honour the trust, the trustee receives payoff T while the truster receives S. The payoffs are ordered so that that the trustee abuses trust if the truster places it (*i.e.*, T > R > P) and the truster prefers not to place trust rather than find his or her trust abused (*i.e.*, R > P > S). The letters T, R, P, S stand for Temptation, Reward, Punishment, and Sucker's Payoff, respectively, and are commonly used to denote payoffs in the Prisoners' Dilemma game. Unless stated otherwise, payoffs are assumed to correspond to players' utilities.

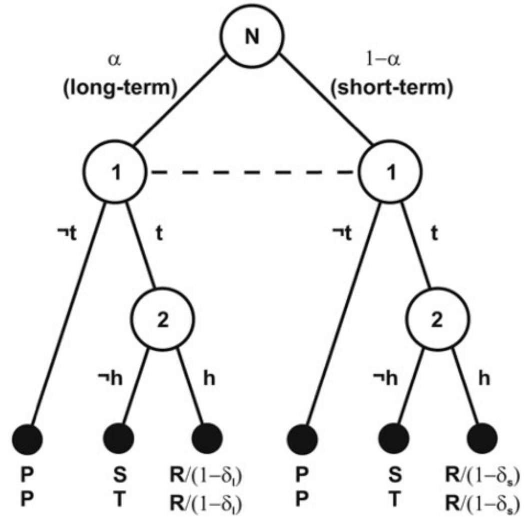

**Figure 2** In the trust game with incomplete information, Nature (N) moves first and determines the trustee's type to be either long term or short term with probability $\alpha$ or $1-\alpha$, respectively. The probability $\alpha$ is common knowledge and the fact that the truster does not know who is who, is denoted by the dashed line. If the truster places trust, the long-term trustee honours trust, whereas a short-term trustee does not do so with certainty. In the first case, the truster receives payoff $R/(1-\delta_l)$; in the second case, the truster's payoff is S. The truster prefers placing trust if the trustee is a long-term type over not placing trust at all and is most reluctant to trust a short-term trustee (*i.e.*, $R/(1-\delta_l) > P > S$).

one-time abuse of trust [*i.e.*, $T > R/(1-\delta_s) > P$]. The long-term and the short-term trustee differ, if their discount factors differ, such that

$$\delta_l > \frac{T-R}{T} > \delta_s \tag{1}$$

Note that an implicit assumption of our model is that an interaction between a truster and a trustee ends if the truster does not place trust or the trustee abuses placed trust.[3] This assumption implies that only the long-term trustee will be deterred from abusing a truster's trust, as he or she would otherwise forgo the higher future benefits from a cooperative repeated interaction with the truster. The short-term trustee's potential future benefits from a repeated interaction with the truster are too small for him or her to resist the temptation of abusing the truster's trust right away.

## Trust Game with Incomplete Information

A truster's uncertainty about a trustee's type can be accounted for in the trust game with incomplete information (Figure 2). Given the probability $\alpha$ to meet a long-term trustee and the payoff structure, a truster only trusts if the expected payoff from trusting is higher than the payoff from not doing so. That is, if

$$\alpha R \frac{1}{1-\delta_l} + (1-\alpha)S > P \tag{2}$$

After solving Equation 2 for $\alpha$, it can be shown that a truster will abstain from placing trust if $\alpha$ is less than the threshold value $\alpha^*$,[4] where

$$\alpha^* = \frac{(P-S)(1-\delta_l)}{R-S(1-\delta_l)} \tag{3}$$

Under these conditions, the truster and the long-term trustee could attain a more beneficial outcome if the trustee were able to communicate his or her type credibly.

## Signalling Trust Game

The model can be extended so that the trustee can initially choose whether to send a signal at cost $c$ (Figure 3). For the truster to interpret the trustee's type, the signal must be type-separating. Then, a separating equilibrium can emerge in which the long-term trustee sends a signal, a short-term trustee does not (*i.e.*, type-separating behaviour), and the truster places trust

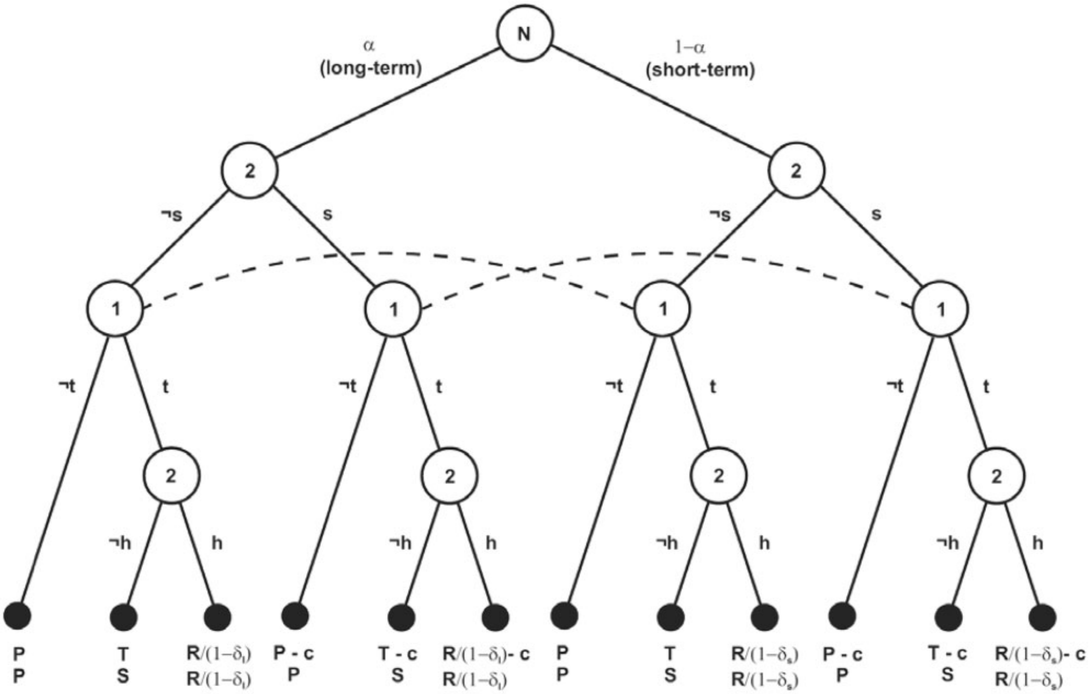

**Figure 3** In the signalling trust game too, first Nature (N) determines the trustee's type, but before the truster decides whether or not to place trust, the trustee decides whether or not to send a signal. The signal is associated with a cost c for the trustee and this cost is incurred first and irrespectively of what the truster and the trustee do thereafter. Before the truster decides whether or not to place trust, they observe the trustee's signalling decision.

only if a signal has been sent by the trustee. The signal is type-separating if the long-term trustee can afford to send it while the short-term trustee cannot. That is, if

$$R\frac{1}{1-\delta_l} - P > c > T - P \qquad (4)$$

If $\alpha < \alpha^*$, the separating equilibrium is collectively more beneficial than the equilibrium without a signalling opportunity. That is, if signalling is not possible, trusters will abstain from placing trust and all will receive payoff $P$. In the former case, however, trusters' expected payoff is $\alpha R/(1-\delta_l) + (1-\alpha)P$ and trustees receive $\alpha[R/(1-\delta_l) - c] + (1-\alpha)P$. If $\alpha > \alpha^*$, a separating equilibrium can also emerge but does not improve collective gains (*i.e.*, the sum of trusters' and trustees' expected payoffs). Without a signalling opportunity, trusters will always place trust and their expected payoff is $\alpha R/(1-\delta_l) + (1-\alpha)S$, with trustees receiving $\alpha[R/(1-\delta_l)] + (1-\alpha)T$. It can be shown that these collective gains are always larger than the collective gains in the separating equilibrium, if $c > T - P$. In other words, if $\alpha > \alpha^*$, a pooling equilibrium, where both trustee types do not send signals (*i.e.*, type-pooling behaviour) and trusters always place trust, is collectively more beneficial.

## Hypotheses

The signalling trust game has three stages at which trusters or trustees have to make decisions (see Figure 3). Hypotheses are derived from the game theoretic model for all three decision stages: (i) the trustee's signalling decision conditional on the trustee's type, (ii) the truster's decision conditional on the trustee's signalling opportunity and the trustee's signal, and (iii) the trustee's decision to cooperate or to defect conditional on the trustee's type.

Note that the game theoretic model devised above implies a binary signalling decision, whereas in our experiments, we implement continuous signals, where trustees can decide how much they want to spend on a signal. However, irrespective of whether the signal is binary or continuous, in a separating equilibrium, the long-term trustee sends a signal at a cost $c$, where $R/(1-\delta_l) - P > c > T - P$, whereas the short-term trustee does not send a signal, and the truster places trust after signalling and withholds trust otherwise. We hypothesize that if $\alpha < \alpha^*$, the strategies in the separating equilibrium are the basis for subjects' behaviour in our experiments. Thus, if at all, long-term trustees will be more likely to deviate from equilibrium signalling from 'above' (*i.e.*, $c' < T - P$), whereas short-term trustees will be more likely to deviate from 'below' (*i.e.*, $c' > 0$). Hence, if trusters observe out-of-equilibrium signalling, they may still infer that signals on which higher amounts have been spent are more likely to be from long-term types than from short-term types.

Our first set of hypotheses suggests that long-term trustees spend higher amounts on signals than short-term trustees, but only if trusters do not have an incentive to unconditionally place trust (*i.e.*, if $\alpha < \alpha^*$). If $\alpha > \alpha^*$, we expect to observe type-pooling signalling behaviour.

*H1a*: If $\alpha < \alpha^*$, the amount spent on a signal by a long-term trustee will be higher than the amount spent by a short-term trustee.

*H1b*: If $\alpha > \alpha^*$, the amount spent on a signal by a long-term trustee will be the same as the amount spent by a short-term trustee.

*H1c*: The short-term trustee's amount spent on a signal will not differ according to the relation between $\alpha$ and $\alpha^*$.

Our second and central set of hypotheses concerns the proportion of trusters placing trust conditional on the signalling opportunity and the amount spent on a signal by the trustee. We hypothesize that a signalling opportunity will increase the level of trust, but only if the trusters do not have an incentive to unconditionally place trust (*i.e.*, if $\alpha < \alpha^*$), and that trusters' decisions to trust will be affected by the amount spent on the signal by the trustees.

*H2a*: If $\alpha < \alpha^*$, the proportion of trusters placing trust will be higher in a game with a signalling opportunity than in a game without a signalling opportunity.

*H2b*: If $\alpha > \alpha^*$, the proportion of trusters placing trust will be the same in a game with a signalling opportunity as in a game without a signalling opportunity.

*H2c*: Irrespective of the signalling opportunity, the proportion of trusters placing trust will be lowest if $\alpha < \alpha^*$ and highest if $\alpha > \alpha^*$.

*H2d*: The higher the amount spent on a signal by the trustee is, the higher will be the proportion of trusters placing trust.

Finally, we expect long-term trustees to be more trustworthy than short-term trustees:

*H3*: For both conditions $\alpha < \alpha^*$ and $\alpha > \alpha^*$, the proportion of long-term trustees honouring trust will be higher than the proportion of short-term trustees honouring trust.

## Experimental Procedure and Design

In total, we conducted four similar computerized experiments; these took place in the United Kingdom (Oxford, Experiment 1), Russia (Nizhny Novgorod, Experiment 2), and Switzerland (Zurich, Experiments 3 and 4) using student populations there.[5] The findings were surprisingly robust across countries and experiments. The most extended experiment was conducted in Oxford, and we report the design and the results from this experiment here. We report the design and the results from the other three experiments in the online Supplementary Data 2.

In each experimental session, subjects were randomly assigned to be a truster or a trustee and kept their roles throughout the session. In accordance with the trust game, a truster could choose between action $t$ (trust) and action $\neg t$ (no trust), and only if the truster decided in favour of action $t$ could the trustee choose between action $h$ (honour trust) and action $\neg h$ (abuse trust). Payoffs were set to $T = 20\,\text{MU}$, $R = 10\,\text{MU}$, $P = 5\,\text{MU}$, and $S = -10\,\text{MU}$, where $5\,\text{MU}$ (monetary units) corresponded to £1. Trusters and trustees were endowed with $20\,\text{MU}$ each at the beginning of every interaction. Moreover, trustees could be either short term or long term. The discount factors of short-term and long-term trustees were set to $\delta_s = 0$ and $\delta_l = 0.667$, respectively, and the expected possible number of games per interaction with either trustee type was derived from the discount factors. Subjects were told that an interaction with a short-term trustee consisted of only one game and that an interaction with a long-term trustee consisted of three repeated games on average, but that a game could be repeated only if the truster chose $t$ and the long-term trustee chose $h$. Subjects were not told that in an interaction with a long-term trustee the minimum possible number of games was two and the maximum was four and that the possible number of games of two, three, or four had been predetermined by the experimenters.

While in every interaction a trustee was told whether he or she was short term or long term, a truster only knew that the proportion of interactions with a long-term trustee was $\alpha$ and the proportion of interactions with a short-term trustee was $1 - \alpha$. In each experimental session, $\alpha$ was either low ($\alpha = 0.25$) or high ($\alpha = 0.50$) and therefore was either below or above the threshold $\alpha^*$ of 0.375 (see Equation 3).

Moreover, subjects were told that the experiment consisted of two parts and that they would only learn at the end of the first part what the second part was about. The signalling opportunity was either introduced in the first part and was absent in the second part, or it was the other way around. In the part with a signalling opportunity, at the beginning of every interaction, a trustee could give up part of his or her endowment to send a card to the truster. The trustee could choose from 11 different cards, with each card having a different price, ranging from 0 to 20 MU. The truster received the card the trustee had chosen and was informed about the expenditure the trustee had made. Next, the truster could make his or her choice in the first trust game of the interaction. The threshold ($c^* = T - P$) above which a card was affordable by a long-term trustee only was 15 MU (see Equation 4). In the part without a signalling opportunity, a trustee could not send any cards to the truster. Table 1 describes our experimental design in more detail.

Throughout the experiment, subjects interacted with alternating participants. Although every truster was paired with every trustee at most three times, subjects did not know who they were paired with in any of the interactions and were paired with the same participants at irregular intervals.

The subjects were 172 students and university employees (53% female, average age: 29.0, $sd = 10.3$). At the beginning of each session, subjects were given written instructions on paper. The instructions were neutrally phrased and are reproduced in the online Supplementary Data 2. After they had read these to themselves, they took a quiz about the instructions, and the correct answers were explained orally to all subjects. Subjects were allowed to refer to the instruction sheets throughout the experiment. Then, the experiment was conducted. After the experiment, participants were asked to fill out a questionnaire and received the money they had earned in the experiment. An experimental session lasted for about 90 min and subjects earned £15 on average (≈€17 or sFr. 25), including a participation fee of £4.

# Results

We present the results for each stage of the signalling trust game separately. The test statistics referred to in this section, and the graphs in Figure 4 are based on the regression model estimations presented in Table A1 in the appendix. Statistical significance is set at the 5% level for two-sided tests, and we account for the repeated measures obtained on the same subject by estimating cluster-robust standard errors.

## Trustees' Signalling Behaviour

Figure 4a shows that long-term trustees spend significantly higher amounts on signals than short-term trustees, which is the case in both the low-alpha (H1a: $F_{1,85} = 37.48$, $P < 0.001$) and the high-alpha condition (H1b: $F_{1,85} = 8.23$, $P = 0.005$). While we expected long-term trustees to send costlier signals to distinguish themselves from short-term trustees in the low-alpha condition, we did not expect such behaviour in the high-alpha condition. However, the difference in the low-alpha condition is significantly larger than in the high-alpha condition ($F_{1,85} = 11.05$, $P = 0.001$), indicating that signalling behaviour is more type-separating in the former than in the latter. In line with hypothesis H1c, the average amount short-term trustees spend on signals in the low-alpha condition does not differ from the amount they spend in the high-alpha condition (H1c: $F_{1,85} = 0.42$, $P = 0.518$). But how do trusters act on the signals that trustees send? In particular, do trusters place more trust the more trustees spend on signals? Moreover, does a signalling opportunity increase the overall level of trust as compared with a condition where no such opportunity exists? We try to answer these questions next.

**Table 1.** Experimental Design

| | Proportion of long-term trustees | |
|---|---|---|
| | Low-alpha ($\alpha = 0.25$) | High-alpha ($\alpha = 0.50$) |
| No signalling first; signalling second | Sequence of 2 × 12 interactions | Sequence of 2 × 10 interactions |
| Signalling first; no signalling second | Sequence of 2 × 12 interactions | Sequence of 2 × 10 interactions |

*We implemented a 2 (low-alpha vs. high-alpha) × 2 (signalling versus no signalling) factorial design. In the low-alpha condition, $\alpha = 0.25$ was below the critical threshold value $\alpha^* = 0.375$. In the high-alpha condition, $\alpha = 0.5$ was above the critical threshold value. Moreover, the experiment consisted of two parts, each part lasting for 12 or 10 interactions, where $\alpha = 0.25$ (i.e., 4/12) or $\alpha = 0.5$ (i.e., 5/10), respectively. We varied α between subject and whether the trustee had a signalling opportunity within subject. That is, each experimental session was conducted with α being either low or high, and the signalling opportunity was introduced either in the first or in the second part. In four experimental sessions, the signalling opportunity was introduced in the first part and in the other four sessions in the second part.*

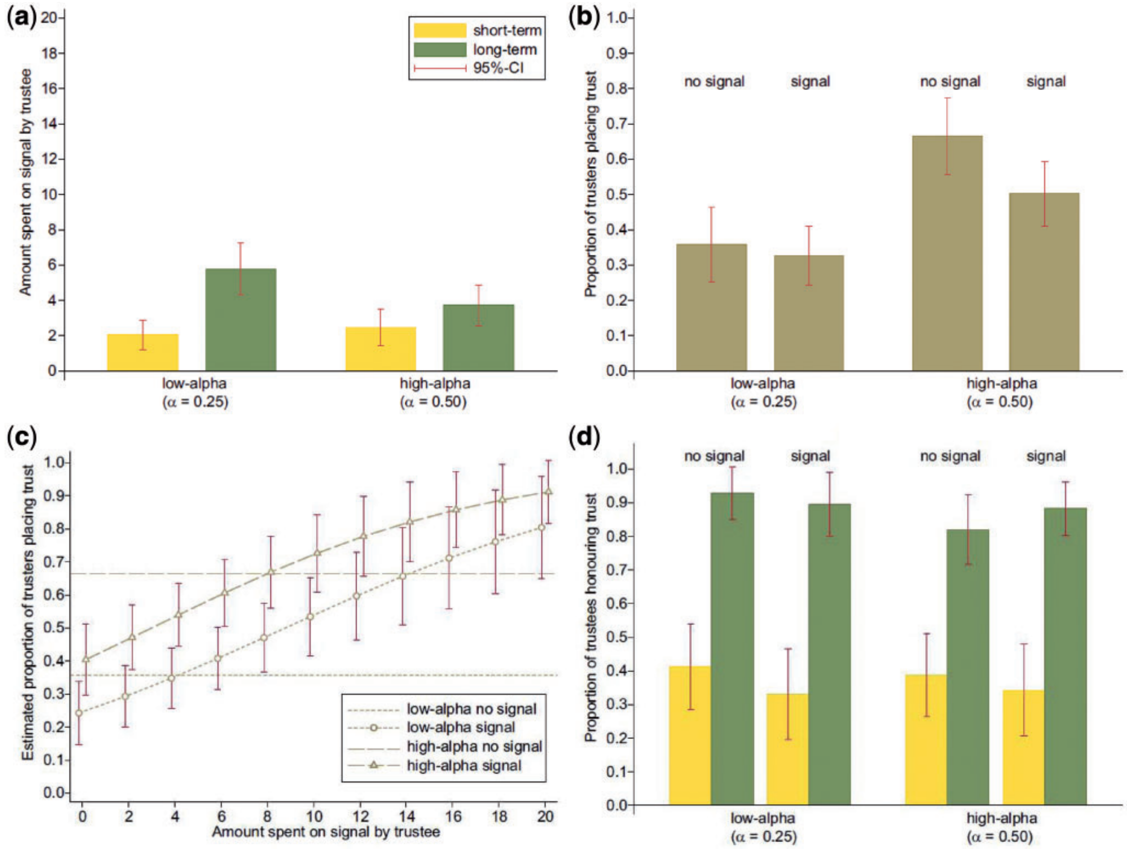

**Figure 4** Main results of Experiment 1. Figures (a) through (d) are based on the regression model estimations that are listed and explained in Table A1 in the appendix. Panel (a) shows the average amounts spent by short-term and long-term trustees on signals in the low-alpha and the high-alpha condition. Long-term trustees spend significantly higher amounts on signals than short-term trustees, but this difference is significantly smaller in the high-alpha condition. Panel (b) shows the proportion of trusters' trusting decisions in the low-alpha and the high-alpha condition, both with and without a signalling opportunity. The overall level of trust does not increase with a signalling opportunity and even decreases in the high-alpha condition. However, the plot in panel (c) shows that the higher the amount a trustee spends on the signal, the higher is the proportion of trusters' trusting decisions in both the low-alpha and the high-alpha condition. Panel (d) shows short-term and long-term trustees' levels of trustworthiness. Long-term trustees are substantially more trustworthy than short-term trustees across all experimental conditions.

## Trusters' Trust

Figure 4b shows the proportion of trusters placing trust in the low-alpha and the high-alpha condition both with and without a signalling opportunity. Contrary to our expectations, a signalling opportunity does not increase the overall level of trust in the low-alpha condition (H2a: $\chi^2_{(1)} = 0.45$, $P = 0.501$) and even decreases trust in the high-alpha condition (H2b: $\chi^2_{(1)} = 8.99$, $P = 0.003$). In line with hypothesis H2c, we observe a significantly higher level of trust in the high-alpha condition than in the low-alpha condition (H2c: $\chi^2_{(1)} = 13.74$, $P < 0.001$). Evidently, without a signalling opportunity, trusters respond to the incentives in the two alpha conditions as expected, but once a signalling opportunity is introduced, their behaviour leads to unexpected outcomes. However, Figure 4c reveals that, in line with our expectations, the more a trustee spends on the signal, the higher is the proportion of trusters placing trust (H2d: $z = 4.02$, $P < 0.001$). So how is it that the overall level of trust does not increase or even decreases once trustees are given the opportunity to communicate their type?

The answer to this question becomes apparent with a closer look at Figure 4c. In both alpha conditions, if trustees send cheap signals (*e.g.*, 0 MU), the proportion of trusters placing trust falls below the proportion observed without a signalling opportunity. In other words, cheap signals induce distrust and trustees have to pay to be trusted to the same extent as they are trusted in the condition without signalling. Based on a rough extrapolation, one can imagine what the outcome would be if the level of unconditional trust decreased by another 30 percentage points while trusters' responsiveness to trustees' signals (*i.e.*, the slope coefficient of the amount spent on the signal) stayed the same. In such a case, it is likely that the overall level of trust induced by trustees' signals would be higher than the trust level without a signalling opportunity. In fact, the low-alpha condition was supposed to induce the level of unconditional trust at zero, but the subjects who participated in our experiment did not behave in the way our game theoretic model would imply.

## Trustees' Trustworthiness

In line with our expectations, we find that the proportion of long-term trustees honouring trust is significantly higher than the proportion of short-term trustees honouring trust (H3: $\chi^2_{(1)} = 23.16$, $P < 0.001$). Across all experimental conditions, the level of long-term trustees' trustworthiness is between 82% and 93%, while it is between 33% and 41% for short-term trustees. The latter clearly indicates that a considerable proportion of short-term trustees act in a non-selfish way, resulting in placed trust being honoured more often than we had expected. It is therefore plausible that trusters placed trust because they acted on their non-zero beliefs about short-term trustees' trustworthiness and these beliefs were reinforced by their positive experience several rounds into the experiment.

Based on these findings, we may conclude that, if the level of unconditional trust is high, the introduction of a signalling opportunity does not further increase that level and could even decrease it, as sending cheap signals will cause distrust. However, at least in the low-alpha condition, the signalling opportunity may still increase collective gains as it allows trusters to distinguish long-term and short-term trustees by their signalling behaviour and to engage in mutually beneficial repeated interactions with long-term trustees only.

## Collective Gains

Tables A2 and A3 in the appendix list trusters' and trustees' expected payoffs, respectively, in each of the four experimental conditions, and explain how these numbers are calculated. It turns out that in the low-alpha condition with a signalling opportunity, the proportion of trust towards long-term trustees is 34% and significantly different from the marginal distribution of 25% ($\chi^2_{(1)} = 10.21$, $P = 0.001$). This difference is small (51–50%) and insignificant in the high-alpha condition. Also, it appears that in the low-alpha condition, an average truster does slightly better with a signalling opportunity (5.48 MU) than without (4.95 MU), and that trusting conditional on trustees' signalling behaviour pays off ($6.46 > 5$ MU) while trusting unconditionally does not ($4.87 < 5$ MU). In the high-alpha condition, trusting clearly pays off, even without a signalling opportunity ($10.05 > 5$ MU), and introducing a signalling opportunity increases these benefits ($11.28 > 5$ MU). Although an average trustee earns considerably more than an average truster and a long-term trustee earns more than a short-term trustee, introducing a signalling opportunity reduces trustees' benefits by 3.20 MU in the low-alpha condition and by 6.15 MU in the high-alpha condition. This roughly corresponds to what trustees have to spend on signals in the low-alpha and the high-alpha condition, respectively, to be trusted to the same extent as they are trusted without a signalling opportunity (see Figure 4c). Hence, in our experiment, trusters win and trustees lose while collective gains decrease when a signalling opportunity is introduced.

## Switzerland and Russia

We conducted three further experiments in Switzerland and Russia. Unlike in the UK experiment described here, in these other experiments, we only implemented the low-alpha condition (*i.e.*, $\alpha < \alpha^*$) and compared the conditions with and without a signalling opportunity in a between-subject design. Results from these experiments are very much in line with the findings presented and discussed here. In particular, the hypothesis of a trust-enhancing effect of the signalling opportunity was not confirmed in any of the three experiments. Furthermore, the short-term and long-term trustees' signalling behaviour and trustworthiness were in accordance with our hypotheses. The results of these experiments are presented and discussed in more detail in the online Supplementary Data 2.

# Discussion and Conclusions

Signalling theory captures social interactions in terms of the information conveyed in the behaviour or other observable characteristics of interacting agents. It starts from the assumption that agents are endowed with different sets of information about each other's preferences and constraints. While agents' preferences and constraints are not observable, agents would benefit from accurate information about them. Posner (1998, 2000) suggests that seemingly irrational behaviour or social norms emerge because they help to distinguish agents who prefer to engage in repeated cooperative interactions (long-term types) from agents with immediate non-cooperative incentives (short-term types).

In this article, we formalized Posner's idea in a signalling trust game with incomplete information and tested hypotheses derived from our model in a series of laboratory experiments. The variables used in the experiments were the probability of encountering a long-term trustee (low-alpha vs. high-alpha) and trustees' possibility of sending costly signals (no signalling opportunity vs. signalling opportunity). Subjects were randomly assigned to be trusters or trustees and trustees could be either short term or long term. Trusters only knew the probability of being paired with either trustee type. Interactions with long-term trustees consisted of repeated games; these consisted of three games on average. Interactions with short-term trustees were one-shot games.

Consistent with our theoretical predictions, long-term trustees spent significantly higher amounts on signals than short-term trustees (H1a), the proportion of trusters' placing trust was higher in the high-alpha condition than in the low-alpha condition (H2c), the proportion increased with the amount a trustee spent on a signal (H2d), and long-term trustees honoured trust significantly more often than short-term trustees (H3). Although the experimental data are in accordance with most of our hypotheses, the core hypothesis of our signalling model (H2b) was not corroborated. Contrary to our expectations, a signalling opportunity did not enhance the level of trust in the low-alpha condition and even decreased it in the high-alpha condition.

In our analysis of the experimental data, we show that a likely cause for the lack of support for H2b is the considerable proportion of short-term trustees honouring trust, presumably because of non-selfish motives. If the *a priori* level of trustworthiness is high (*i.e.*, $>\alpha^*$), trusters have a real incentive to unconditionally place trust in the low-alpha condition without a signalling opportunity. Thus, a signalling opportunity that is supposed to increase the trust level from zero to $\alpha = 0.25$, once introduced, may not exhibit the trust-enhancing effect relative to the actually existing level of 36%. What is more, in the high-alpha condition, where the level of unconditional trust is supposed to be 100%, the introduction of a signalling opportunity may cause the trust level to decrease if type-separating rather than type-pooling signalling behaviour emerges. That is, if long-term trustees start distinguishing themselves from short-term trustees by sending costly signals, cheap signals may induce distrust, as they will be expected to come from short-term types and, as a result, the level of trust will decrease. The fact that we find type-separating signalling behaviour in both alpha conditions, albeit to a significantly lower extent in the high-alpha condition, suggests that type-separating signalling can emerge more gradually and that the degree of type distinction it exhibits may depend on the actual level of uncertainty about trustees' trustworthiness. In other words, the less trusters are willing to place trust unconditionally, the more long-term trustees are willing to spend on costly signals to distinguish themselves from short-term types. Finally, we observed that the introduction of a signalling opportunity in both the low-alpha and the high-alpha condition slightly increased trusters' expected payoffs while it considerably decreased trustees' expected payoffs, resulting in a net loss in collective gains. However, it remains an open question as to how far a signalling system in which trustees' signals are broadcast to more than one truster (*e.g.*, through the internet or at public events) would increase collective gains without increasing the level of trust.

What do these findings tell us about Posner's signalling theory in general and our signalling model in particular? Obviously, Posner's signalling theory has inspired our thinking about social cooperation and has led us to suggest a formalized version of it in this article. Such formalization will, we believe, ultimately allow us to explain seemingly irrational (signalling) behaviour in human social interaction and to predict under what conditions it is likely to emerge. However, before we can get to the latter, we need to reconsider the way in which we implemented our experiments to test the predictions from our model.

Note first that most of our hypotheses are supported by our findings and the results that are not in line with our hypotheses can be plausibly explained within the scope of the model. This may sound contradictory, but in fact it is not. The signalling trust game that we devised for this article assumes—but does not require—actors to be self-interested. Thus, the fact that in our experiment a high proportion of short-term trustees honoured trust for non-selfish reasons changes the conditions under which we put our model to a test, but it does not invalidate our model. However, there are several aspects of the implementation of our experiments that could be improved in future research. In particular, the *a priori* level of trustworthiness of short-term trustees could be decreased by a decrease in the monetary incentives for honouring trust in the one-shot trust game. At the same time, this would reduce trusters' beliefs about short-term trustees' trustworthiness and bring down *a priori* trust to a level that has the potential to be increased by a signalling opportunity. However, the problem with this approach is that a change in subjects' monetary incentives also changes the parameter values of our model, which still assumes self-interested actors. In other words, the model makes new predictions as a result of the change in payoffs, whereas our aim in changing the payoffs is to obtain the conditions under which the old predictions can be tested. The easiest way out of this dilemma would be to use deception, *i.e.*, let subjects interact with trusters and trustees who are programmed to behave in a rational and selfish way while making them believe that they are interacting with human subjects. This approach, however, is controversial, mostly because it is believed that experimental subjects, once they learn that deception is being used, will also believe that it is being used even if it is not. Ironically, these subjects' wrong beliefs may then distort the experimental conditions one attempts to create to test a particular hypothesis.[6]

In our attempt to implement the *experimentum crucis* without using deception, we have learned something interesting about actual human behaviour and the functioning of a signalling system. In particular, we have learned that if *a priori* levels of trust and trustworthiness are high, introducing a signalling opportunity that is meant to distinguish long-term and short-term types may have a counterproductive effect. Therefore, we suppose that the predictions of Posner's theory crucially depend on the *a priori* levels of trust and trustworthiness and that a signalling opportunity will increase individual and collective gains if these levels are low. This tells us that the existing levels of trust and trustworthiness are important contextual variables that must be accounted for if one wants to study the emergence of signalling behaviour in the field. Finally, the fact that we do not find major differences across three countries tells us that cultural differences may be unimportant in explaining subjects' behaviour in experimental signalling trust games.

## Notes

1. Both interpretations are in line with Posner's theoretical ideas (see *e.g.*, Rasmusen, 2001, chapter 5). However, here we adhere to the former interpretation, as it allows us to take a sociological perspective and to manipulate the discount factor experimentally.
2. The present value of the reward from repeated cooperation is $R/(1-\delta) = R + \delta R + \delta^2 R + \delta^3 R + \ldots$
3. This assumption is similar to assuming that in an infinitely repeated trust game, the truster and the trustee use a so-called trigger strategy, which starts defecting forever once the other party abstains from cooperation (see *e.g.*, Osborne, 2004, chapter 14). In our model, we additionally assume that only mutual cooperation can go on forever, but not mutual defection. We make this assumption because it makes our model more comprehensible. In Footnote 1 in the online Supplementary Data 1, we show that the qualitative model predictions are unaffected by the latter assumption.
4. Coleman (1990: 99) introduces a similar decision threshold in his chapter on trust relations. His formalization of the trust problem, however, does not account for the strategic nature of the decision situation (see also Voss, 1998).
5. All experiments were programmed and conducted with the software z-Tree (Fischbacher, 2007). The data from all experiments will be made available on the authors' home pages.
6. In fact, we conducted an experiment in which subjects in the role of trusters interacted with computer-simulated trustees in a long sequence of interactions. However, we did not use deception; we informed subjects that they would interact with simulated trustees and not with a real person. In this experiment, short-term trustees were programmed to always abuse the truster's trust, and long-term trustees were programmed to always honour that trust. Moreover, in the condition with a signalling opportunity, trustees' signalling behaviour was noisy but obviously correlated with their type. Although it took trusters some time to learn to respond optimally, the results of this experiment are consistent with our conjecture that the introduction of a signalling opportunity increases trust if the *a priori* level of trustworthiness is low (see Przepiorka, 2009: 79–86).

## Acknowledgements

We are grateful for helpful comments made by Friedel Bolle, Vincent Buskens, Vincenz Frey, Diego Gambetta, Debra Hevenstone, Ben Jann, Siegwart Lindenberg, Chris Snijders, Manuela Vieth, and five anonymous reviewers. We are indebted to Nina Ivashinenko and Alexander Soldatkin for making it possible to conduct the experiment in Russia. We would like to thank Stefan Wehrli and Matthias Naef from DeSciL, the experimental laboratory at ETH Zurich, for their help in conducting the experiments in Switzerland. One experiment was conducted at the Centre for Experimental Social Sciences (CESS) at the University of Oxford, thanks to the indispensable support of Luis Miller and Paloma Ubeda.

## Funding

This work was supported by the Swiss National Science Foundation [grant number 100015_124877].

## References

Ahn, T. K. and Esarey, J. (2008). A Dynamic model of generalized social trust. *Journal of Theoretical Politics*, **20**, 151–180.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Bacharach, M. and Gambetta, D. (2001). Trust in signs. In Cook, K. S. (Ed.), *Trust in Society*. New York: Russell Sage Foundation, pp. 148–184.

Bliege Bird, R. and Smith, E. A. (2005). Signaling theory, strategic interaction, and symbolic capital. *Current Anthropology*, **46**, 221–248.

Bolle, F. and Kaehler, J. (2007). Introducing a signaling institution: an experimental investigation. *Journal of Institutional and Theoretical Economics*, **163**, 428–447.

Braun, N. (1992). Altruismus, Moralität und Vertrauen. *Analyse und Kritik*, **14**, 177–186.

Buskens, V. (1999). *Social Networks and Trust*. Utrecht: ICS dissertation.

Buskens, V. and Raub, W. (2002). Embedded trust: control and learning. In Lawler, E. J. and Thye, S. R. (Eds.), *Group Cohesion, Trust and Solidarity*. Vol. 19, Advances in Group Processes. Amsterdam: JAI, pp. 167–202.

Camerer, C. and Weigelt, K. (1988). Experimental test of a sequential equilibrium reputation model. *Econometrica*, **56**, 1–36.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, **74**, 1579–1601.

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, MA: The Belknap Press of Harvard University Press.

Cook, K. S. and Hardin, R. (2001). Norms of cooperativeness and networks of trust. In Hechter, M. and Opp, K.-D. (Eds.), *Social Norms*. New York: Russell Sage Foundation, pp. 327–347.

Dasgupta, P. (1988). Trust as a commodity. In Gambetta, D. (Ed.), *Trust: Making and Breaking Cooperative Relations*. Oxford: Basil Blackwell, pp. 49–72.

Diekmann, A. and Przepiorka, W. (2010). Soziale Normen als Signale. Der Beitrag der Signaling-Theorie. In Albert, G. and Sigmund, S. (Eds.), *Soziologische Theorie kontrovers. Sonderheft 50 der Kölner Zeitschrift für Soziologie und Sozialpsychologie.* Wiesbaden: VS Verlag, pp. 220–237.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, **10**, 171–178.

Gambetta, D. (2009). Signaling. In Hedström, P. and Bearman, P. (Eds.), *The Oxford Handbook of Analytical Sociology*. Oxford: Oxford University Press, pp. 168–194.

Hardin, R. (1996). Trustworthiness. *Ethics*, **107**, 26–42.

Johnstone, R. A. (1997). The evolution of animal signals. In Krebs, J. R. and Davies, N. B. (Eds.), *Behavioral Ecology: An Evolutionary Approach*. Oxford: Blackwell, pp. 155–178.

Kreps, D. (1990). Corporate culture and economic theory. In Alt, J. E. and Shepsle, K. A. (Eds.), *Perspectives on Positive Political Economy*. Cambridge, MA: Cambridge University Press, pp. 90–143.

McAdams, R. H. (2001). Signaling discount rates: law, norms, and economic methodology. *Yale Law Journal*, **110**, 625–689.

Nelson, P. (1974). Advertising as information. *Journal of Political Economy*, **82**, 729–754.

Osborne, M. J. (2004). *An Introduction to Game Theory*. Oxford: Oxford University Press.

Posner, E. A. (1998). Symbols, signals, and social norms in politics and the law. *Journal of Legal Studies*, **27**, 765–798.

Posner, E. A. (2000). *Law and Social Norms*. Cambridge, MA: Harvard University Press.

Przepiorka, W. (2009). Reputation and signals of trustworthiness in social interactions. Zurich: ETH Zurich (Diss. No 18649).

Rasmusen, E. (2001). *Games and Information: An Introduction to Game Theory*. Malden, MA: Blackwell Publishing.

Raub, W. (2004). Hostage posting as a mechanism of trust: binding, compensation, and signalling. *Rationality and Society*, **16**, 319–365.

Riegelsberger, J., Sasse, M. A. and McCarthy, J. D. (2005). The mechanics of trust: a framework for research and design. *International Journal of Human-Computer Studies*, **62**, 381–422.

Snijders, C. (1996). *Trust and Commitments*. Amsterdam: Thela Thesis.

Snijders, C. and Buskens, V. (2001). How to convince someone that you can be trusted? The role of 'hostages'. *Journal of Mathematical Sociology*, **25**, 355–383.

Spence, M. A. (1973). Job market signalling. *Quarterly Journal of Economics*, **87**, 355–374.

Vieth, M. (2009). *Commitments and Reciprocity. Experimental Studies on Obligation, Indignation, and Self-Consistency*. Utrecht: ICS dissertation.

Voss, T. (1998). Vertrauen in modernen Gesellschaften. Eine spieltheoretische Analyse. In Metze, R., Mühler, K. and Opp, K.-D. (Eds.), *Der Transformationsprozess: Analysen und Befunde aus dem Leipziger Institut für Soziologie*. Leipzig: Leipziger Universitätsverlag, pp. 91–129.

# Appendix

**Table A1.** Regression model estimations

| | (a) Trustee spent on signal (in MU) OLS | | (b) Truster placed trust (0/1) Logit 1 | | (c) Truster placed trust (0/1) Logit 2 | | (d) Trustee honoured trust (0/1) Logit 3 | |
|---|---|---|---|---|---|---|---|---|
| Trustee type [short-term (ref.), long-term] | | | | | | | | |
| Long-term | 3.736*** | (0.610) | | | | | 2.918*** | (0.606) |
| Alpha condition [$\alpha = 0.25$ (ref.), $\alpha = 0.50$] | | | | | | | | |
| $\alpha = 0.50$ | 0.428 | (0.659) | 1.268*** | (0.342) | 1.268*** | (0.342) | −0.103 | (0.375) |
| Signalling opportunity [no signal (ref.), signal] | | | | | | | | |
| Signal | | | −0.138 | (0.205) | −0.555* | (0.236) | −0.353 | (0.258) |
| Amount spent on signal (0–20 MU) | | | | | | | | |
| Amount | | | | | 0.128*** | (0.032) | | |
| Two-way interactions | | | | | | | | |
| Long-term × $\alpha = 0.50$ | −2.490** | (0.749) | | | | | −0.944 | (0.718) |
| Long-term × signal | | | | | | | −0.072 | (0.601) |
| $\alpha = 0.50$ × Signal | | | −0.539 | (0.305) | −0.518 | (0.351) | 0.159 | (0.364) |
| $\alpha = 0.50$ × amount | | | | | 0.009 | (0.047) | | |
| Three-way interaction | | | | | | | | |
| Long-term × $\alpha = 0.50$ × Signal | | | | | | | 0.768 | (0.783) |
| const. | 2.047*** | (0.410) | −0.582* | (0.236) | −0.582* | (0.236) | −0.353 | (0.266) |
| $N_1$ | 946 | | 1892 | | 1892 | | 856 | |
| $N_2$ | 86 | | 86 | | 86 | | 86 | |
| adj. $R^2$ | 0.07 | | | | | | | |
| pseudo $R^2$ | | | 0.05 | | 0.08 | | 0.20 | |
| $\chi^2$ | | | 20.43 | | 46.03 | | 80.75 | |

*Notes: The table lists coefficient estimates from OLS and logistic regression models with cluster-robust standard errors in parentheses (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$ for two-sided tests). The dependent variable in Model (a) is the amount spent on a signal by a trustee. The binary dependent variable in Models (b) and (c) is 1 if the truster placed trust and the binary dependent variable in Model (d) is 1 if the trustee honoured trust. The graphs shown in Figures 4a–d are based on these respective model estimations. While Figures 4a–d are based on variants of the respective models estimated without a constant and a full set of interaction terms, Figure 4c shows predicted trust levels for different amounts spent on the signal by a trustee in each experimental condition. These predictions are based on Model (c).*

**Table A2.** Trusters' expected payoffs

| Low-alpha | Without signalling opportunity | | | | | With signalling opportunity | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | t (0.36) | | | | ¬t (0.64) | t (0.33) | | | | ¬t (0.67) |
| | Short-term (0.77) | | Long-term (0.23) | | – | Short-term (0.66) | | Long-term (0.34) | | – |
| | h (0.41) | ¬h (0.59) | h (0.93) | ¬h (0.07) | – | h (0.33) | ¬h (0.67) | h (0.89) | ¬h (0.11) | – |
| Payoff | 10 | −10 | 30 | −10 | 5 | 10 | −10 | 30 | −10 | 5 |
| Expected | 1.14 | −1.64 | 2.31 | −0.06 | 3.20 | 0.72 | −1.46 | 3.00 | −0.12 | 3.35 |
| | 4.95 (total); 4.87 (t); 5.00 (¬t) | | | | | 5.48 (total); 6.46 (t); 5.00 (¬t) | | | | |

| High-alpha | Without signalling opportunity | | | | | With signalling opportunity | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | t (0.67) | | | | ¬t (0.33) | t (0.50) | | | | ¬t (0.50) |
| | Short-term (0.51) | | Long-term (0.49) | | – | Short-term (0.49) | | Long-term (0.51) | | – |
| | h (0.39) | ¬h (0.61) | h (0.82) | ¬h (0.18) | – | h (0.34) | ¬h (0.66) | h (0.88) | ¬h (0.12) | – |
| Payoff | 10 | −10 | 30 | −10 | 5 | 10 | −10 | 30 | −10 | 5 |
| Expected | 1.33 | −2.08 | 8.08 | −0.59 | 1.65 | 0.83 | −1.62 | 6.73 | −0.31 | 2.50 |
| | 8.38 (total); 10.05 (t); 5.00 (¬t) | | | | | 8.14 (total); 11.28 (t); 5.00 (¬t) | | | | |

*Notes: The table lists the proportions of trusters' trusting decisions (t) and the proportions of short-term and long-term trustees' decisions to honour placed trust (h) in each of the four experimental conditions. Based on these proportions and the payoffs of the trust game, trusters' (total) expected payoffs as well as their expected payoffs from placing trust (t) and not placing trust (¬t) are calculated. For example, the expected value of the path [t, short-term, ¬h] in the low-alpha condition with signalling opportunity is 0.33 × 0.66 × 0.67 × ( − 10) = −1.46. The sums of the corresponding values are listed at the bottom of each of the four sub-tables. Note that it is assumed that an interaction with a long-term trustee who honours placed trust in the first game lasts for three games and therefore yields a payoff of 30 MU for the truster.*

**Table A3.** Trustees' expected payoffs

| Low-alpha | Without signalling opportunity | | | | With signalling opportunity | | | |
|---|---|---|---|---|---|---|---|---|
| | Short-term (0.75) | | Long-term (0.25) | | Short-term (0.75) | | Long-term (0.25) | |
| | t′ (0.37) | ¬t′ (0.63) | t′ (0.33) | ¬t′ (0.67) | t′ (0.29) | ¬t′ (0.71) | t′ (0.44) | ¬t′ (0.56) |
| Payoff | 20 | 5 | 30 | 5 | 20 | 5 | 30 | 5 |
| Expected | 5.55 | 2.36 | 2.48 | 0.84 | 4.35 | 2.66 | 3.30 | 0.70 |
| | 11.23 (total); 10.55 (s); 13.25 (l) | | | | 8.03 (total); 7.30 (s); 10.22 (l) | | | |

| High-alpha | Without signalling opportunity | | | | With signalling opportunity | | | |
|---|---|---|---|---|---|---|---|---|
| | Short-term (0.50) | | Long-term (0.50) | | Short-term (0.50) | | Long-term (0.50) | |
| | t′ (0.68) | ¬t′ (0.32) | t′ (0.65) | ¬t′ (0.35) | t′ (0.49) | ¬t′ (0.51) | t′ (0.52) | ¬t′ (0.48) |
| Payoff | 20 | 5 | 30 | 5 | 20 | 5 | 30 | 5 |
| Expected | 6.80 | 0.80 | 9.75 | 0.88 | 4.90 | 1.28 | 7.80 | 1.20 |
| | 18.23 (total); 15.20 (s); 21.25 (l) | | | | 12.08 (total); 9.88 (s); 14.28 (l) | | | |

*Notes: The table lists the proportions of short-term and long-term trustees that have been trusted (t′) or not trusted (¬t′) in each of the four experimental conditions. Based on these proportions and the payoffs of the trust game, trustees' (total) expected payoffs as well as the expected payoffs of short-term (s) and long-term trustees (l) are calculated. For example, the expected value of the path [short-term, ¬t′] in the low-alpha condition with signalling opportunity is 0.75 × 0.71 × (5) = 2.66. The sums of the corresponding values are listed at the bottom of each of the four sub-tables. Note that in the conditions with a signalling opportunity, the average amounts spent on signals have been subtracted. The average amounts spent on signals are 2.98 MU (total), 2.05 MU (s), and 5.78 MU (l) in the low-alpha condition and 3.10 MU (total), 2.47 MU (s), and 3.72 MU (l) in the high-alpha condition (see Figure 4a). Finally, note that it is assumed that a short-term trustee does not honour trust and therefore earns 20 MU, whereas a long-term trustee honours trust in three consecutive games and therefore earns 30 MU.*