



It takes two to cheat: An experiment on *derived* trust

Maria Bigoni^{a,*}, Stefania Bortolotti^a, Marco Casari^a, Diego Gambetta^{b,c}

^a Economics Department, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy

^b European University Institute, Fiesole, Italy

^c Nuffield College – Oxford University, UK



ARTICLE INFO

Article history:

Received 8 January 2013

Accepted 19 August 2013

Available online 11 September 2013

JEL classification:

C92

C72

D03

Keywords:

Trust game

Coordination

Inequality aversion

Reciprocity

Collective trust

ABSTRACT

Social life offers innumerable instances in which trust decisions involve multiple agents. Of particular interest is the case when a breach of trust is not profitable if carried out in isolation, but requires an agreement among agents. In such situations the pattern of behaviors is richer than in dyadic games, because even opportunistic trustees who would breach trust when alone may act trustworthily based on what they believe to be the predominant course of action. Anticipating this, trusters may be more inclined to trust. We dub these motivations derived trustworthiness and derived trust. To capture them, we design a “Collective Trust Game” and study it by means of a laboratory experiment. We report that overall levels of trustworthiness are almost thirty percentage points higher when derived motivations are present, and this generates also higher levels of trust. In our set-up, the effects of derived trustworthiness are comparable in size to positive reciprocity, and more important than concerns for equality.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The standard conceptualization of trust rests on a dyadic relation between a truster and a trustee. The archetype, which often social scientists have in mind, is an elementary market transaction between, say, a lender and a borrower: if the lender lends, the borrower must decide whether to return the money or pocket it. The lender most pressing question is whether the borrower will be driven by his most basic self-interest or he will be able to resist that temptation and behave trustworthily.

Trust relations, however, offer innumerable instances in which trustees are not acting in isolation, but, especially if they are untrustworthy, may have to take into account the actions of other trustees before deciding whether or not to breach trust themselves. Think for instance of two bystanders, strangers to each other, observing someone dropping his wallet and observing each other observing the event. They could nod to each other in agreement, pick up the wallet and share its valuable content. If only one of them is so inclined, in order to pull off his act he needs at least the acquiescence of the other. But if even only one of them might be honest and want to call the attention of the owner, the dishonest bystander is in a dilemma: will any indication of his true preferences cost him at least a reproach and shame? Is he not better off just pretending to go along with an honest decision?

Less trivially, imagine a hospital in which the management entrusts the purchase of medical supplies to a number of its employees, on the explicit understanding that they will look for the cheapest deals. Each employee knows that by secretly

* Corresponding author. Tel.: +39 051 209 8122; fax: +39 051 0544 522.

E-mail addresses: maria.bigoni@unibo.it (M. Bigoni), stefania.bortolotti@unibo.it (S. Bortolotti), marco.casari@unibo.it (M. Casari), diego.gambetta@eui.eu (D. Gambetta).

agreeing with his colleagues to buy supplies from a particular dealer they could all gain-for, sheltered from competition, the dealer can charge a higher price and share the ill-gotten gain with the corrupt employees. Whether or not each employee will breach the hospital trust and pursue such a corrupt agreement depends on his probity of course, but lacking this, it depends on what he expects the other employees' behavior would be, and on the consequences in case his corrupt behavior is detected. In a situation such as this even opportunistic trustees – who would breach trust when deciding in isolation – may act as if they were trustworthy if they believe trustworthiness is the predominant course of action, and if the consequences for non conformance are bad enough.

Communities in which trustworthiness is believed to be the norm might generate more of trustworthiness than one could gauge simply taking into account the trustworthiness generated by pro-social virtues. In many settings, trustworthiness would have a positive externality, act in other words as a trust *multiplier*. It would spread its beneficial bacteria to those who, as pure self-interested maximizers, would otherwise be immune to it.¹ In order to capture this multiplicative effect, which cannot emerge within dyadic trust relationships, we have designed a game with three players – one truster and two trustees – which we dub the “Collective Trust Game.”

Our goal is to measure how the behavior of an opportunistic trustee changes when a norm of trustworthiness can be enforced, as compared to situations in which it cannot be. We study experimentally two variants of the Collective Trust Game. In the *Baseline* case, there is no punishment for the untrustworthy trustee, while in the *Coordination* case there is a monetary sanction for a trustee who breaches trust alone. Consider the hospital example, and suppose that most employees are self-interested and opportunistic. They would rather collude with each other and with the supplier, in order to make money at the expenses of the hospital. Even if it is known that the general custom is to be honest, but there are no consequences for those who try to breach trust alone, the opportunistic employees would try their luck and attempt to collude. However, if the sanctions for those caught red-handed are harsh enough, nobody would try to instigate corruption because of the expectation that a disapproving colleague may inform on them. We dub the choice to act honestly (i.e., repay trust) “*derived trustworthiness*” if it is self-regarding and depends on the belief that other trustees will also be trustworthy. Derived trustworthiness is captured in the *Coordination* variant of the Collective Trust Game where there is an explicit sanction for mis-coordination, while it is absent in the *Baseline* variant of the game.² Notice that trustworthiness and self-interest are typically misaligned, but *only* if one expects others to be trustworthy they can work in unison. Symmetrically, the hospital manager can entrust also self-interested employees if he anticipates that the untrustworthy employees are reluctant to reveal their type. We label the choice to entrust the trustees “*derived trust*” if it hinges on the belief that there exists some derived trustworthiness in the population. A comparison of the two variants of the game allows us to isolate the derived component of trust and trustworthiness.

To assess the relative importance of derived trustworthiness and of other typical determinants of trustworthiness – such as positive reciprocity and concerns for equality – also present in dyadic situations, we run two additional treatments: *Baseline-Passive* and *Coordination-Passive*. In both variants of the game, a robot takes a random decision on behalf of the truster. There is a participant with the role of truster but makes no choices, although she fully bears the payoff consequences of other's actions. In these *Passive* treatments, trustees' decisions should not be motivated by reciprocity because in case of an action favorable to them, it is the result of chance and not of the kindness of the truster.

We report three main findings. First, trustworthiness is a thirty percentage points higher in treatments in which derived motives are present (i.e., *Coordination*). Second, trusters grasp the strength of the derived motivation to be trustworthy and are considerably more trusting when penalties for mis-coordination among trustees are present and trust is more profitable in these treatments. Third, with our parametrization derived trustworthiness is comparable in size to the fraction of trustworthiness motivated by positive reciprocity, and larger than the fraction originating from concerns for equality.

The structure of the paper is as follows. [Section 2](#) discusses the existing literature and [Section 3](#) presents the Collective Trust Game. [Section 4](#) describes the experimental design and procedures. [Section 5](#) discusses the main results of the experiment and [Section 6](#) concludes.

2. Literature review

We say that a person trusts another if she acts on the expectation that the trustee(s) will do X, which is a particular task,³ when both truster and trustee know that two conditions are obtained: (i) if the trustee fails to do X, the truster would have done better to act otherwise; but if the trustee does X then the truster is better off than if she had acted otherwise; (ii) if trusted, the trustee has the opportunity to pursue a selfish reason not to do X – in other words, the truster voluntarily puts

¹ Anderlini and Terlizzese (2012) propose a theoretical model which captures that same sort of multiplicative effect by assuming that trustees pay a ‘cost of cheating’, which is the larger, the smaller the mass of agents who cheat.

² Our definition of derived trustworthiness only considers extrinsic motivations that can induce a self-regarding profit-maximizer agent to act trustworthily. Intrinsic motivations to conform to others' actions, that could be present in the *Baseline* variant as well, are not included in the present definition.

³ The notion of trust with which we work combines elements proposed by Coleman (1990), Fehr (2009)'s concise interpretation of Coleman (1990)'s, and Bacharach and Gambetta (2001). We think of trust as task-specific, largely because the motivations of the trustee to be trustworthy can be task specific: what makes one not jump a red light, steal from a neighbor who gave him the keys to feed the cat, or take care of a bicycle or a car lent to him may differ, and one may trust someone in one case and not another.

herself in a vulnerable position.⁴ In many but not all situations a third condition applies, namely also the trustee is (iii) better off being trusted than not being trusted, regardless of whether he is or is not trustworthy.⁵ We say that a trustee is trustworthy if he simply does X when the above conditions apply.

The essence of trust has been captured in laboratory experiments by means of the so-called investment game, also known as trust game (Berg et al., 1995). In such game, a principal (truster, hereafter) receives an endowment E and has to decide how much of this endowment to invest. The amount invested, k , is exogenously multiplied – usually by a factor $\alpha = 3$ – and transferred to an agent (trustee, hereafter). The trustee has to decide the amount, r , to return to the truster. When considering the raw payoffs of the games and not the all-in payoffs, a self-interested profit-maximizer trustee should not return any money and, by backward induction, the truster should not invest. In contrast, trust and trustworthiness have commonly been observed in the lab and this result is robust to a number of structural variations (for a review of the literature, see Johnson and Mislin, 2011; Camerer, 2003). To subvert the theoretical prediction and to trust, a truster must believe that the trustee possesses some trust-warranting property that stops him from yielding to the raw payoffs and, all things considered, makes him behave trustworthily. Some all-in payoffs that induce trustworthiness can express the trustee's long-term self-interest, such as that which is satisfied by building a reputation (Dasgupta, 1988). But the most interesting trust-warranting properties typically induce a deviation from self-interest however defined. There are several trust-warranting properties of this kind: (i) positive reciprocity; (ii) (unconditional) other-regarding preferences (Cox, 2004; McCabe et al., 2003); (iii) trust responsiveness (Bacharach et al., 2007); and (iv) guilt aversion (Charness and Dufwenberg, 2006).⁶ Cox (2004) and McCabe et al. (2003) report that positive reciprocity accounts for a large part of the observed trustworthiness, even though (unconditional) other-regarding preferences are non-negligible. Similarly, the truster may trust out of one of the following motivations: (i) (unconditional) other-regarding preferences; (ii) risk attitudes; and the (iii) beliefs about opponent's trustworthiness (see, for instance, Cox, 2004; Eckel and Wilson, 2004; Sapienza et al., 2007; Schechter, 2007; Houser et al., 2010a).

A number of scholars have enriched the classic framework, by moving from dyadic to multi-players trust settings (i.e., multiple trusters or trustees). While fitting the most cited definitions of trust (for instance, see Coleman, 1990; Fehr, 2009; Bacharach and Gambetta, 2001), multi-players settings also bring about new trust-warranting properties overlooked when one employs a dyadic framework. Cassar and Rigdon (2011) provide evidence that trust and trustworthiness are inherently comparative and what matters the most for trustworthiness is the relative, rather than the nominal, amount received. Mittone and Ploner (2009) study the role of social effects, such as peer pressure and social spillovers, on trustworthiness in a modified investment game with multiple trustees, and report that peer pressure increases reciprocity. Social effects are also studied by Regner and Riener (2011) who provide evidence that when trustees move sequentially, the second imitates the first, but only if that is in her best interest. Sheremeta and Zhang (2012) study a sequential 3-player trust game where two trusters decide sequentially – any amount sent for one to the other is tripled – and find that communication between the second truster and the trustee greatly improves trust at any node of the game. Finally, McEvily et al. (2006) show that trusters generalize previous experience with one member of a group to the entire group, even when group membership is created via minimal group paradigm.

Extending this line of research, our work investigates the role of payoff-interdependence among trustees by means of a novel game we dub the Collective Trust Game. The Collective Trust Game captures a situation involving one truster and two trustees with interdependent payoffs. This introduces two elements absent in a dyadic setting. First, the truster's decisions rely not only on his beliefs about the dispositions of the single trustees, but also on his beliefs about trustees' interactions – derived trust. Second, trustees' payoffs are interdependent and the response to a trusting decision also depends on what is believed to be the predominant course of action in the population – derived trustworthiness. In our novel set-up trustees choose simultaneously and without communication. The interdependence between trustees hence operates only through the interaction of the decisions after the choice is made. This game bears resemblance to the Collective Resistance Game employed in Political Science (Weingast, 1997; Cason and Mui, 2007), which differs from the Collective Trust Game in that the first mover (“leader”) has the chance to increase her own payoff by reducing (rather than increasing) the social welfare.

3. Collective trust game

In the experiment participants repeatedly face a Collective Trust Game with changing opponents and with random switches between the role of truster and trustee. We are interested in understanding the many instances of social life in which people encounter others and face a sequence of opportunities to trust and be trusted. In this settings there can be a “contagion effect” of trustworthiness in a population. We employ two variants of the Collective Trust Game, *Coordination* and *Baseline* (Fig. 1). In both variants, the truster and the trustees face binary decisions. The truster has to choose whether to trust (action “Send”) or not (action “Keep”). If the truster chooses Keep, the two trustees have no choice to make and everyone gets 20. If the truster chooses Send, each of the trustees chooses between “Breach” and “Return.”⁷ Trust increases

⁴ See also Rousseau et al. (1998, p. 395): “...intention to accept vulnerability based upon positive expectations of the intentions or behavior of others”.

⁵ This is not true of trust situations in which the trustee is merely saddled with a burden – as when he is asked, say, to look after his neighbor's cat, a case in which only (i) and (ii) obtain.

⁶ As reported in Bacharach et al. (2007, p. 350), trust responsiveness is a “tendency to fulfill trust because you believe that it has been placed on you”.

⁷ In the experiment, the actions had neutral labels. See the Instructions in Appendix D.

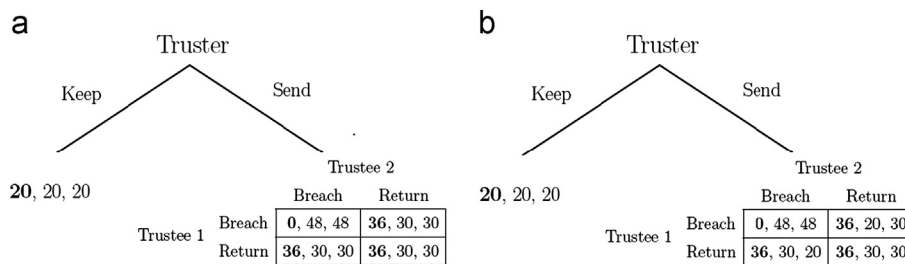


Fig. 1. Collective trust game. (a) Baseline, (b) coordination.

Notes: The first payoff (in bold) indicates the truster's earnings, while the second and the third report the earnings of Trustee 1 and Trustee 2, respectively. Trustees 1 and 2 decide simultaneously. Earnings are in tokens; at the end of each session, participants received 1 Euro for every 40 tokens.

aggregate surplus in any case, but it pays off for the truster only if at least one trustee returns trust. In this case, the truster earns 36, while he gets nothing if all trustees choose Breach. Both variants satisfy the definitions of trust in Section 2 but differ in trustees' payoffs, in particular, in the way they are interdependent.

In *Coordination*, when they are trusted, the trustees earn 30 (48) each if they both Return (Breach). Nevertheless, players face a penalty for mis-coordination: a trustee who plays Breach, while the other plays Return, earns 20 (instead of 30). This situation resembles the case in which a self-interested employee is stigmatized and ostracized by his honest co-workers after imprudently revealing his bad intentions. Notice that the sanction to breach trust alone is relatively mild: indeed, the earnings for the untrustworthy player are the same as if he were not trusted in the first place and no further sanction is applied.

The subgame played by the trustees is similar to the stag-hunt and has three Nash equilibria: a payoff-dominant equilibrium and a payoff-dominated equilibrium in pure strategies, and an equilibrium in mixed strategy. In the payoff-dominant equilibrium, both trustees choose Breach and get a payoff equal to 48, each. In the payoff-dominated equilibrium, both trustees choose Return and get a payoff of 30. In the mixed strategy equilibrium, each trustee plays Return with probability 9/14. For the truster, trust always pays off unless trustees coordinate on the payoff-dominant equilibrium (Breach, Breach). The break-even point in expected terms is when each trustee returns with an independent probability of 1/3. This value yields for the truster a payoffs of 0 with probability 4/9 and a payoff of 36 with probability 5/9.

In other words, the *Coordination* variant of the Collective Trust Game admits multiple subgame perfect Nash equilibria, and the outcome depends on the equilibrium selection criterion in use in the reference group. If the predominant course of action among trustees prescribes to take advantage of people whenever possible then the truster is worse off trusting than not trusting. If instead the prevalent line of conduct either inclines the trustees towards the mixed strategy equilibrium, or prescribes to be trustworthy, then the truster is better off trusting.

Baseline differs from *Coordination*, because in the former there is no penalty for mis-coordination of actions among trustees (Fig. 1); in the hospital example, there is no social sanction for the employee caught red-handed by his co-workers. *Baseline* is the closest to the standard trust game with two agents.⁸ As in *Coordination*, the subgame between trustees has two Pareto-ranked Nash equilibria in pure strategies and one in mixed strategies. In one equilibrium both trustees choose Breach and individual payoffs are 48, while in the other equilibrium both trustees choose Return and individual payoffs are 30.

A further difference emerges between *Baseline* and *Coordination*: in *Baseline*, the equilibrium (Send, Return, Return) does not survive elimination of weakly dominated strategies. Besides, even though *Baseline* nominally admits multiple subgame perfect Nash equilibria, there exists a unique evolutionary stable equilibrium.⁹ As before, trust pays off in expected terms if each trustee chooses Return with an independent probability of 1/3. Hence, trust does not pay if the trustees coordinate on the evolutionary stable equilibrium.

This design allows us to capture what we labeled the derived components of trust and trustworthiness. A comparison of *Baseline* and *Coordination* allows us to isolate the derived component in trust and trustworthiness from other determinants and to assess the relative importance of each. The working definition we use to defined the derived motives is as follows:

Derived trustworthiness. A trustee's choice to Return is driven by "derived trustworthiness" if it is self-regarding and depends on the belief that other trustees will also choose Return.

Derived trust. A truster's choice of Send is motivated by "derived trust" if it hinges on the belief that there exists some derived trustworthiness in the population.

Notice that the definition of derived trustworthiness does not include herd behavior while it well suits the definitions of trust proposed in the literature. In *Baseline* (and of course in the dyadic trust game) there is no extrinsic motivations for

⁸ Although playing a dyadic trust game might seem an attractive control, the comparison with a Collective Trust Game is not straightforward because the differences can originate from differences both in payoffs and in game structure. We could keep constant across these games either the individual payoffs (simply dropping the second trustee) or the wealth multiplier, but not both at the same time.

⁹ The equilibrium (Send, Return, Return) does not withstand a refinement based on evolutionary arguments, see proof in Appendix A.

Table 1
Experimental treatments.

Treatment	Baseline active	Coordination active	Baseline passive	Coordination passive
Stage game	Baseline	Coordination	Baseline	Coordination
Truster	Human	Human	Robot	Robot
Motives for trustworthiness				
Inequality aversion	✓	✓	✓	✓
Reciprocity	✓	✓		
Derived trustworthiness		✓		✓
Session	03/31;04/04	03/30;04/01	04/15	05/13
Subjects	60	57	30	27
Independent matching sets	4	4	2	2

Notes: Sessions conducted in 2011. In *Baseline-Passive*, because of a technical problem, the average frequency of trust was 53%, which is slightly different from the average frequency of 47% observed in *Baseline-Active*. There were two independent matching sets in each session, see main text for details.

self-interested profit-maximizer participants to choose Return if he expects that the other trustee will choose Return, hence neither derived trust nor derived trustworthiness plays a role.¹⁰

4. Experimental design

We implemented a two by two factorial between-subjects design (Table 1). One dimension of variation concerns the way trustees' payoffs are interdependent in the Collective Trust Game. In the *Coordination* treatments there is a monetary penalty for mis-coordination, while no such a penalty is present in the *Baseline* treatments (Fig. 1). The other dimension concerns whether the truster's choice to Send or Keep is intentional (*Active* treatments) or is made by the computer through a random draw on behalf of the participant (*Passive* treatments). In all treatments, the decision is payoff relevant for the truster.¹¹ In the *Passive* treatments, the random choice to Send or Keep followed a period-specific probability distribution that replicated the empirical distribution of choices made in the same period of the corresponding *Active* treatment (pooling all sessions).

The four treatments elicit a different mix of trust-warranting motives. In the *Baseline-Active* treatment, trustees can choose Return for two main reasons: positive reciprocity or aversion to inequality. Hence, either they want to reciprocate a kind and intentional action of the truster, or they care about equality and want to reduce the differences in earnings between themselves and the truster.

In the *Coordination-Active* treatment an additional motive is present: derived trustworthiness. This motivation stems from the need to avoid the loss carried by mis-coordination; that is, self-interested trustees may choose Return simply because this maximizes their expected payoff. A comparison between these treatments allows us to address our two main questions: does trustworthiness have a *derived* component? does trust have a *derived* component?

Finally, in the *Passive* treatments reciprocal motives are removed on the assumption that one cannot be grateful to a truster when the decision is made by a machine. In the *Baseline-Passive* treatment equality-concerns are the only possible explanation for choosing Return. Recall that in the *Passive* treatments, while trust decisions are made at random, there are in the room participants in stand-by who earn the payoffs of the truster. Hence, the two trustees' decisions in a group impact their earnings as well as the earnings of one other person in the same way as in the *Active* treatments. In the *Coordination-Passive* treatment also derived motives come into place due to coordination concerns. A comparison between our four treatments allows us to disentangle the impact of derived trustworthiness, from the impact of positive reciprocity and of aversion to inequality.

Each session included a choice over lotteries (part one), three dictator games (part two), and a repeated Collective Trust Game (part three), in this order. We gave no feedback about choices and earnings of parts one and two until the end of the session, to minimize possible spill-over effects on subsequent decisions (see, for instance, Houser et al., 2010b). We will now describe part three and then move to the description of part one and two.

Each participant played 30 repetitions of the Collective Trust Game, in different roles and with changing opponents. Before every period, random groups of $N=3$ were formed (strangers matching protocol) and then the roles of truster and trustee were assigned at random. Changing roles can facilitate learning and help spreading norms of trust within the population.¹² In order to increase the statistical power, we partitioned participants in each session into two matching sets, in

¹⁰ In our definition, derived trustworthiness is not motivated by preferences for conformity or such like, but is only driven by self-interest. In addition, even if there were a preference for conformity, it should apply both to the *Baseline* and to the *Coordination* treatment, and it should lower the utility of both players whenever they mis-coordinate. Hence, it is not obvious whether such a preference would induce more or less trustworthiness. We believe that there is no derived trustworthiness in the *Baseline* treatment even if social preferences may generate multiple equilibria.

¹¹ Cox (2004)'s treatment C has a design similar to our *Passive* treatments.

¹² Some studies pointed out that playing both roles may reduce both the levels of trust and trustworthiness (Burks et al., 2003; Johnson and Mislin, 2011). However, in our *Coordination* treatment, trustworthiness levels are high enough to make trust profitable. Changing roles may hence promote rather than hinder both trust and trustworthiness.

Table 2

Lottery task: choose one option.

Lottery	High payoff (orange ball)	Low payoff (white ball)	CRRA coefficient	Frequency of choices (%)
Lottery 1	17.5	17.5	> 3.64	13
Lottery 2	22.5	15.0	3.46–1.16	24
Lottery 3	27.5	12.5	1.16–0.70	18
Lottery 4	32.5	10.0	0.70–0.50	27
Lottery 5	37.5	7.5	0.50–0	16
Lottery 6	44.0	1.0	< 0	2

Notes: The Coefficient of Relative Risk Aversion (CRRA) was not shown to participants. Payoffs for the lottery task were presented in Euros.

a way that a participant in one set never met a participant from another set. Participants interacted repeatedly within the same population. Hence, it was relatively easy for them to form an accurate belief about the predominant course of action in their matching set. The interaction was anonymous, hence decisions could not be based on the identity of who was in the group. By design, individuals could not build a reputation, as we did not provide information on individual identity or history of play.

In the *Active* treatment, if the truster chose Send, then the trustees had to simultaneously choose either Breach or Return. If the truster chose Keep, then the trustees had no decision to make.¹³ In the *Passive* treatments, in each period the trustees observed a draw from an urn containing blue and yellow balls. The draw selected the first move done by the computer for the passive person who was sitting in for the truster. Trustees had to make a decision only in case a blue ball was drawn. Subjects did not receive any information about the composition of the urn or how the composition was determined. However, we made clear to the participants that the composition was determined in advance and was unaffected by their choices during the game. The urn composition changed from period to period so as to reproduce the same distribution of Send actions observed in the *Active* treatments.

At the end of each period, participants could observe the choice made by every player in their group as well as their individual period earnings. They received no information about subjects outside their group. Before periods 1 and 30, we elicited participants' beliefs about trust and then about trustworthiness in the population. There was no payment associated with belief decisions to avoid possible interferences with choices in the trust game (Croson, 2000).¹⁴

In the first part of each session, participants had to choose one lottery from the list of six options described in Table 2. Each lottery had a low and a high payoff outcome that always occurred with a 50% chance (Eckel and Grossman, 2002, 2008). Only two participants per session received a payment for the choice over lotteries.¹⁵

In the second part of each session, participants faced three dictator games: DG1, DG2, and DG3. In each task a dictator allocated a given amount of tokens between herself and two others. In DG1 she chose an option that ranged between allocation E – which granted 160 tokens to the dictator and 160 tokens to each of the other two persons – and allocation W (160, 130, 340). In DG2 the allocations ranged between S (190, 40, 250) and W, while in DG3 between S and E (more details in Appendix B). These tasks provided an estimation of other-regarding preferences at the individual level, which are employed as controls in the regression analysis. The instructions were distributed before each part and were read out aloud by the experimenter (Appendix D).

The experiment involved 174 participants, divided into 6 sessions (Table 1) and was conducted at the Bologna Laboratory for Experiments in Social Sciences (BLESS). Subjects were mostly students at the University of Bologna, recruited through ORSEE (Greiner, 2004). About 53% of the participants were male; nobody took part in more than one session. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007). Upon arrival, participants were randomly assigned to a cubicle to avoid eye contact, and no communication was allowed during the experiment. A quiz to ensure full understanding was administered before the third part; everyone had to correctly answer all questions before proceeding. Before payments, participants answered a questionnaire. The average session lasted about 1 h and 20 min. Subjects were paid privately in cash at the end of the session and earned on an average 23 Euros.

5. Results

This section illustrates the main results and is structured around two parts. In Section 5.1 we study the *Active* treatments and identify the derived motives for trustworthiness and trust (Results 1 and 2). In Section 5.2, we juxtapose data from *Active* and *Passive* treatments to analyze the determinants of trustworthiness. More specifically, we measure the

¹³ If the truster chose Keep, only the choice of the first mover was available since trustees did not enter the second node of the game.

¹⁴ Participants answered the following questions before knowing their role for the period: "Out of every 10 persons in role A [trusters], how many do you expect, on an average, to Keep and how many to Send?" and "Out of every 10 persons in role B [trustees] who face a decision, how many do you expect, on an average, to Breach and how many to Return?"

¹⁵ A volunteer performed two manual draws from a bag with numbered balls to randomly select these participants. An additional manual draw selected the outcome for the lottery chosen by the two participants. To avoid carry over effects, the draws were performed at the end of the session.

Table 3
Trust and trustworthiness by treatment.

Treatment	Baseline active	Coordination active	Baseline passive	Coordination passive
Trust (%)	53	81	–	–
Trustworthiness (%)	36	62	12	50
(Breach, Breach) (%)	41	17	77	24
(Breach, Return) (%)	47	41	22	52
(Return, Return) (%)	12	42	1	24
Earnings	26.4	28.6	25.6	28.4
N. Subjects	60	57	30	27

Notes: Average level of trust and trustworthiness, all periods pooled. Trust is profitable for the truster if at least one of the two trustees decides to play return. We observe a choice between Breach and Return only for groups in which the truster chose Send – i.e., we did not implement the strategy method. Earnings refer to average per-period monetary earnings.

relative weight of derived trustworthiness in the decision to Return as compared to preferences for equality and positive reciprocity (Result 3).

5.1. Derived trust and trustworthiness

The main difference between *Baseline* and *Coordination* lies in the way trustees' choices are interdependent: summary statistics for *Baseline* and *Coordination* are reported in Table 3. This manipulation has a direct effect on trustworthiness and only indirectly affects trust (trusters expect – or experience – different levels of trustworthiness). For this reason, we first analyze trustees' behavior and then study trusters' behavior.

Does trustworthiness have a derived component? In *Coordination-Active*, trustworthiness can be induced by derived motives, which stem from trustees' desire to avoid the cost of mis-coordination. Derived trustworthiness can hence be measured by comparing *Coordination-Active* with *Baseline-Active* where no such motives are in place.

Result 1. Trustworthiness has a large derived component. The relevance and significance of the derived component increases over time.

As shown in Fig. 2, the frequency of the Return actions is always larger when derived motives are in place. Overall, average (median)¹⁶ trustworthiness was 0.36 (0.15) and 0.62 (0.67) in *Baseline-Active* and *Coordination-Active* respectively.¹⁷ Derived trustworthiness seems to play some role already in the first period – 50% of the trustees choose Return in *Coordination-Active* while only 35% do so in *Baseline-Active*. However, initially the difference is not statistically significant.¹⁸ We will next show that the treatment effect is reinforced by experience.

Further support for Result 1 is provided through a series of probit regressions (see Table 4).¹⁹ The dependent variable indicates whether the trustee chose Return (1) or Breach (0) in each period. In Model 1 the only explanatory variable is a dummy for *Coordination-Active* treatment, which is meant to capture the effect of derived trustworthiness. According to results from this regression, the probability that a trustee chooses Return is 44% higher in *Coordination-Active* than in *Baseline-Active*, lending support to Result 1. Variants of this model deliver qualitatively and quantitatively similar results. In Model 2, we include two additional explanatory variables which capture learning effects. In our repeated interaction setting, learning has an upper bound (i.e., the average level of trustworthiness cannot be larger than 1) and tends to be more pronounced in earlier rather than later periods. We therefore opted for the reciprocal of the variable *Period* to account for it. Interactions between $1/Period$ and the treatment dummies are included in the second model of Table 4. The negative coefficients of these interaction terms indicate that trustworthiness increases over time; however, the upward trend is not statistically significant. Result 1 is robust also when controlling for a number of socio-demographic characteristics and for risk aversion and other-regarding preferences as (see Table 4, Models 3 and 4).

We now turn to truster behavior: *Does trust have a derived component?*

¹⁶ Median trustworthiness is computed by considering average individual trustworthiness over the 30 periods. Average trustworthiness is the frequency of Return in the population.

¹⁷ The difference is significant, albeit weakly ($p=0.083$, two-tailed Wilcoxon rank-sum test, $N_1=N_2=4$). Subjects, however, did not correctly anticipate behavioral differences between the two treatments (see Appendix C for a detailed discussion). Indeed, pre-play beliefs are not significantly different across treatments: on an average, participants expected 39% of the trustees to play Return in the *Baseline-Active* and 44% in the *Coordination-Active* treatment ($p=0.278$, two-sample Wilcoxon rank-sum test, $N_1=60$ and $N_2=57$, two-sided). Subjects played 30 repetitions of the stage game and were rematched after each period according to a strangers matching protocol. When performing non-parametric tests, we conservatively use the average at the matching-set level as independent unit of observation. Only for tests in period one, we consider each participant as an independent observation.

¹⁸ A chi-squared test returns $p=0.350$, with $N_1=20$ and $N_2=18$.

¹⁹ Regression models for binary data are adopted throughout the paper since truster and trustees had to make a binary choice (i.e., Send/Keep and Return/Breach). Symbols *, **, and *** denote significance at the 10%, 5% and 1% level, respectively. The paper reports marginal effects from probit regressions; since data are not independent at the individual level over time we included individual random effects. Results are robust to different specifications, such as random effect logit/OLS regressions and probit regressions with errors clustered at the matching set level.

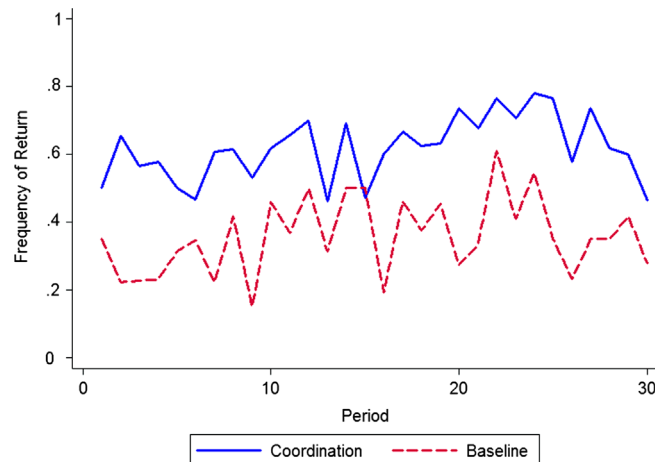


Fig. 2. Trustworthiness across periods: active treatments.

Table 4

Treatment effect on trustworthiness: active treatments.

Dependent variable: Return (1) Breach (0)	Model (1)	Model (2)	Model (3)	Model (4)
Coordination Treatments	0.441*** (0.115)	0.459*** (0.115)	0.344*** (0.091)	0.359*** (0.093)
1/Period \times baseline treatments		–0.033 (0.139)		–0.025 (0.117)
1/Period \times coordination treatments		–0.179 (0.114)		–0.151 (0.097)
Controls	No	No	Yes	Yes
N.obs.	1554	1554	1554	1554

Notes: Marginal effects from probit regressions with individual random effect. Trustees could decide between Breach and Return only if the trustor chose Send. Given the observed frequency of Send, there were 1554 decisions between Breach and Return. In 799 out of 1554 cases (51%), the trustee chose Return over Breach. “Controls” indicates the presence of eleven regressors. Four regressors are built from the incentivized elicitation of individual preferences: *Strongly risk averse* equals 1 for participants choosing the most, or the second-most, risky lottery in Table 2; *Risk neutral/Risk loving* equals 1 for participants choosing the least, or the second-least, risky option; *Strong concerns for group wealth* equals 1 if a participant leans toward option W in both DG1 and DG2 (details in Appendix B); *Strong concerns for equality* equals 1 if a participant leans toward option E in both DG1 and DG3. Seven regressors concern demographics: *Male*; *Age26+* equals 1 if the participant is 26 or older; *Cars2+* equals 1 if the family owns two or more cars; *House of property* equals 1 if the family owns a house; *Siblings* is the number of siblings, plus one (i.e. 1 if no siblings, 4 if three or more); *South-Islands* equals 1 if a participant attended the primary school in the South of Italy, Sardinia or Sicily; *Center* equals 1 if attendance in a central region of Italy, as defined by the Italian Institute of Statistics (ISTAT). Except *Siblings*, all controls are dummy variables.

*** indicates the significance at 1% level.

Result 2. Trust has a large and significant derived component.

The results suggest a positive answer to our second research question, on whether derived trust does matter: the average (median) level of trust over the 30 periods is 0.53 (0.59) in *Baseline-Active* and 0.81 (0.91) in *Coordination-Active*.²⁰ As illustrated in Fig. 3, the initial level of trust is similar across treatments and becomes significantly different only after some experience of the game.²¹ In the initial period trust is indeed statistically indistinguishable across treatments.²² However, after period 1 the gap between the two treatments progressively widens; specifically, a pronounced upward trend emerges when derived motives are present.

Table 5 reports results from four probit regressions providing further support to Result 2. The dependent variable is whether the trustor chose Send (1) or Keep (0) in each period and the explanatory variables are the same as in Table 4. The estimated probability of observing a trustful decision in *Coordination-Active* is between 28 and 36% larger than in *Baseline-Active*, all else held constant. Overall, trustors chose Send more often in *Coordination-Active* than in *Baseline-Active*: the result is highly significant and robust to different specifications and to socio-demographics controls. A strong learning

²⁰ Overall, trust levels are significantly different between the two conditions at 5% level, as revealed by a two-tailed Wilcoxon rank-sum test ($p=0.021$, $N_1=N_2=4$).

²¹ We find that the higher the pre-play belief about others' trustworthiness, the higher the initial level of trust in both treatments (see discussion and Fig. C1 in Appendix C).

²² Average trust level in period 1, is 0.50 and 0.47 in *Baseline-Active* and *Coordination-Active*, respectively; $p=0.869$, chi-squared test, $N_1=20$ and $N_2=19$.

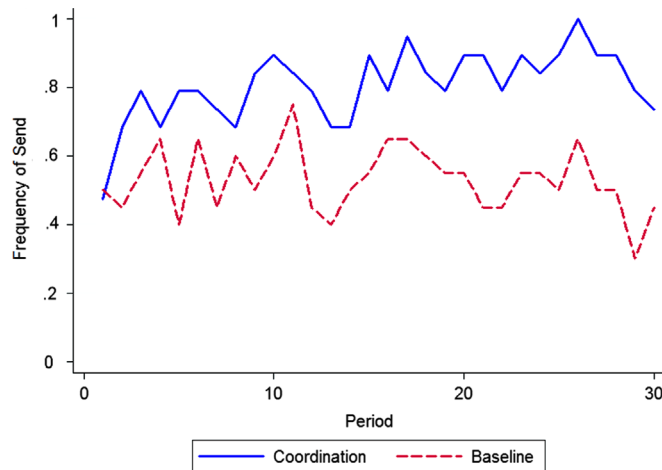


Fig. 3. Trust across periods: active treatments.

Table 5

Treatment effects on trust: active treatments.

Dependent variable: Send (0) Keep (1)	Model 1	Model 2	Model 3	Model 4
Coordination treatments	0.332*** (0.072)	0.361*** (0.071)	0.286*** (0.067)	0.323*** (0.067)
1/Period × baseline treatments		−0.041 (0.088)		−0.043 (0.086)
1/Period × coordination treatments		−0.389*** (0.087)		−0.369*** (0.085)
Controls	No	No	Yes	Yes
N. obs.	1170	1170	1170	1170

Notes: Marginal effects from probit regressions with individual random effect. For Controls see notes to Table 4.

*** indicates the significance at 1% level.

dynamic emerges in *Coordination-Active*: the probability of trusting increases over time, as indicated by the negative coefficient for $1/\text{Period} \times \text{Coordination}$. This evidence suggests that populations starting from similar levels of trust and trustworthiness can converge to different norms if we allow trustworthy behavior to have a positive externality.

As revealed by Table 3, the higher levels of trust and trustworthiness observed in the *Coordination-Active* led subjects to earn on an average more than in the *Baseline-Active* treatment.²³ Hence, the benefit of increased trust outweighed the higher costs of coordination failures.

5.2. Decomposing trustworthiness

What is the relative contribution of derived trustworthiness, positive reciprocity, and concerns for equality? Consider the hospital example, an employee may actually look for the cheapest deal for at least three possible reasons: (i) he believes his colleagues are honest and fears that his mis-conduct will not be profitable because it will be reported – derived trustworthiness; (ii) he is intrinsically honest and wants to repay the trust of his employer – positive reciprocity; (iii) he has a preference for equality and does not want to let someone behind – concerns for equality. In order to disentangle the three motivations behind trustworthiness, we carry out a comparison of Return frequencies across treatments.

Result 3. When decomposing trustworthiness, derived motives appear as a major drive, comparable in size to positive reciprocity, and larger than concerns for equality.

We say that a participant has a concern for equality if he is willing to self-sacrifice to reduce the inequality in his group. Trustees can increase equality by choosing Return (i.e., trustees get 30 each and the truster gets 36) rather than Breach (i.e., if both trustees choose Breach they get 48 each and the truster gets nothing). Whereas concerns for equality are present in

²³ The difference is significant at the 5% level according to a two-tailed Wilcoxon rank-sum test ($p=0.043$, $N_1 = N_2 = 4$). The difference in earnings between the two *Passive* treatments, instead, is not statistically significant.

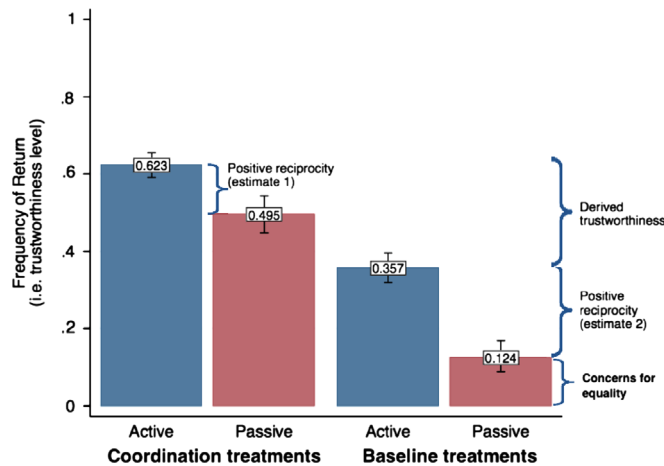


Fig. 4. Possible motivations for being trustworthy: a treatment comparison.

Notes: We illustrate the magnitude of some possible motivations of trustworthy behavior from the trustees' return frequency across treatments. For instance, "derived trustworthiness" is equal to 0.266 and is computed as the difference between the return level 0.623 in the *Coordination-Active* treatment and 0.357 in the *Baseline-Active*. The whiskers at the top of each bar represent exact binomial 95% confidence intervals for the average frequency of return.

all four treatments, there is only one treatment in which neither (i) nor (ii) is present: *Baseline-Passive*. In this treatment, the average Return frequency (0.12, Table 3 and Fig. 4) is sensibly smaller as compared to all other treatments.

Positive reciprocity is defined as the willingness to repay a kind and intentional action of the truster. By design, this attitude is present in *Active* treatments but not in *Passive* treatments, because in the latter a robot decides between Keep or Send. We estimate the impact of positive reciprocity in the range of 12–24% in the frequency of trustworthiness, which results from the comparisons of *Active* and *Passive* treatments. In our design, there are two ways to estimate the relative importance of reciprocal motives for trustworthiness: one is the difference in the *Baseline* treatments (0.36 vs. 0.12) and the other is the same difference in the *Coordination* treatments (0.62 vs. 0.50). In both cases, the impact of reciprocity on trustworthiness is substantial, even though it is smaller in the *Coordination* treatments.

Finally, the derived component of trustworthiness can be isolated by contrasting Return frequencies in *Coordination-Active* and *Baseline-Active* as discussed in the previous section (mean Return frequencies of 0.62 vs. 0.36, respectively). As shown in Fig. 4, the relative importance of derived trustworthiness is at least as large as the impact of positive reciprocity, and much larger than the one of concerns for equality.

To corroborate this result, Table 6 presents a series of probit regressions where the dependent variable is whether the trustee chose Return (1) or Breach (0) in each period. The *Coordination Treatments* dummy captures the derived component of trustworthiness and is positive and highly significant in all estimates. The *Active Treatments* dummy is meant to capture positive reciprocity; while the coefficient is positive, it is noticeable that the marginal effect of positive reciprocity is smaller than the one of derived trustworthiness. Positive reciprocity in *Coordination* is larger than in *Baseline* treatments (Fig. 4). Notice also that different drivers of trustworthiness are not necessarily orthogonal and can interact with each other. For instance, if trustees believe that the truster chose Send only because of strategic reasons, they may be less inclined to reciprocate this action. This crowding out effect between derived trust and positive reciprocity is captured in the regression analysis by the interaction variable *Coordination × Active Treatment*.²⁴

To sum up, the decomposition of trustworthiness in the *Coordination-Active* treatments yields the following quantitative results. The total of 62% of Return choices is attributed for 26 percentage points to strategic motives, 24 percentage points to positive reciprocity, and 12 percentage points to equality concerns. It is quite remarkable that in our multi-players trust game strategic trustworthiness is comparable in size to (if not larger than) positive reciprocity.

6. Discussion and conclusions

We have studied a new trust-warranting property that can be highly relevant in the field and is absent from previous experiments: derived trust and trustworthiness. Even those who have no primary motive to act trustworthily end up doing so if they expect that the other trustee is likely enough to choose to be trustworthy, or if they believe that the other trustee will believe that they will be trustworthy, and so on. Expectations about other trustees' behavior can thus lead self-interested payoff maximizers to be trustworthy. These derived motivations can intervene in addition to and independently of other-regarding preferences.

To study derived trust and trustworthiness we extend the standard conceptualization of trust as a dyadic interaction to a situation with multiple trustees. Our setting, the Collective Trust Game, introduces two novel elements: first, trustees'

²⁴ If different motivations are not mutually exclusive – for instance, when participants are motivated by both derived and reciprocal motives – we risk underestimating the effects of positive reciprocity. It is a possible explanation of the different estimates across treatments.

Table 6
Determinants of trustworthiness: all treatments.

Dependent variable Return(1) Breach(0)	Model 1	Model 2
Coordination treatments	0.639*** (0.073)	0.594*** (0.071)
Active treatments	0.373*** (0.079)	0.344*** (0.080)
Coordination × active treatment	−0.251*** (0.087)	−0.247*** (0.078)
Controls	No	Yes
N.obs.	2276	2276

Notes: Marginal effects from probit regression with individual random effect. Trustees could decide between Breach and Return only if the truster chose Send. Given the observed frequency of Send, there were 2276 decisions between Breach and Return. In 1052 out of 2276 cases (46%), the trustee chose return over breach. For controls see notes to Table 4. Standard errors are reported in parentheses.

*** indicates the significance at 1% level.

payoffs are interdependent and trustees face a coordination issue; second, and because of that, trusters' decisions will also depend on what they believe is the predominant course of action in their communities with regard to trustworthiness, which is related to social norms and conventions. This new setting fits standard definitions of trust.

We report that derived motives strongly affect both trust and trustworthiness. When decomposing trustworthiness, derived motives appear as a major drive, comparable in size to positive reciprocity, and larger than concerns for equality. Moreover, higher levels of trustworthiness also lead to higher levels of trust, which shows that trusters understand the force of the derived motivation.

Our extension is only one among several possible ways to generalize dyadic trust, and the empirical contribution of each motivation reported in this study is of course linked to the parameters we chose, as it is the case for all experimental studies. Multi-agent trust games offers a suitable platform to explore rich set of questions about collective trust. These considerations, however, should not overshadow the evidence of a large and qualitatively different component of trust and trustworthiness that standard experimental investigations of trust cannot capture. In situations where derived trust and trustworthiness are important, differences in the realized levels of trust may be driven by differences in the beliefs over trustees' behavior, not just by individual preferences. Results from our generalization of the dyadic trust game to a game with multiple trustees involved in a coordination problem carry a more optimistic implication than the standard trust game experiment: even self-interested profit-maximizers may act trustworthily in communities in which trustworthiness is believed to be the predominant course of action, and in which non-conformance is punished harshly enough. If the Collective Trust Game represents a frequent type of social situation, it is even conceivable that every day we cross path with potential villains who refrain from forming criminal partnerships and behave trustworthily simply because they do not know each others' true intentions and for fear of incurring a penalty in the case they reveal themselves to the wrong counterpart.

As Schelling (1960, p. 140) wrote: “The bank employee who would like to rob the bank if he could only find an outside collaborator and the bank robber who would like to rob the bank if only he could find an inside accomplice may find it difficult to collaborate because they are unable to identify each other, there being severe penalties in the event that either should declare his intentions to someone who proved not to have identical interests”. We cannot quantify how many breaches of trust fail to materialize because of the multiplier effect of trustworthiness, but it is arguably a very considerable amount. As Hirschman (1984) once suggested, trust is a most peculiar resource that is depleted by not being used.

Acknowledgments

The authors thank Tim Cason, Enrique Fatas, Bendikt Herrmann, Giorgio Negrone, Wojtek Przepiorka, Andrzej Skrzypacz, participants at the 2011 Florence Experimental Economics workshop, the ESA 2011 meetings in Chicago and Luxembourg, 2011 Workshop on Trust and Cultural Evolution in Valencia, and seminar participants at Purdue University and Bologna University. We gratefully acknowledge the financial support from the ERC Starting Grant Strangers 241196. Previous versions of this work circulated under the title “Trustworthy by convention.” The usual disclaimer applies.

Appendix A. Evolutionary stability of trust in the collective trust game

In this appendix, we show that the cooperative strategy (C) that prescribes to Send when in the role of truster, and to Return when in the role of trustee is not evolutionary stable in the *Baseline* variation of the Collective Trust Game, while it is evolutionary stable in the *Coordination* variation.

A.1. Baseline variation

Consider a large population of individuals, all of whom are “programmed” to play the cooperative strategy C . Suppose a small group of “mutants” – programmed to play strategy D – appears in this population. Let the share of mutants in the (post entry) population be $\varepsilon \in (0, 1)$. In every period, sets of three individuals in this population are randomly formed, and roles are also assigned at random, so that each individual has the role of truster with probability $\frac{1}{3}$, and the role of trustee with probability $\frac{2}{3}$. In addition, for each individual the probability that each of the other 2 members of the set will play the mutant strategy D is ε and the probability that she will play the incumbent strategy C is $1 - \varepsilon$. The expected payoff in a period in this bi-morphic population is thus the same as in a match with an individual who plays the mixed strategy $M_1 = \varepsilon D + (1 - \varepsilon)C$. The expected payoff to the cooperative strategy C is $u(C, M_1)$ and that of the mutant strategy $u(D, M_1)$.

According to Weibull (1997), a strategy x is evolutionary stable if for every strategy $y \neq x$ there exists some $\varepsilon_y \in (0, 1)$ such that:

$$u[x, \varepsilon y + (1 - \varepsilon)x] > u[y, \varepsilon y + (1 - \varepsilon)x], \quad \forall \varepsilon \in (0, \varepsilon_y)$$

The intuition is that evolutionary forces will select against a mutant strategy only if that strategy is less “fit” than the incumbent strategy, i.e. if in expectation it generates a lower payoff.

It is straightforward to show that, in the *Baseline* variant of the Collective Trust Game, the cooperative strategy C is not evolutionary stable. In fact, it can be invaded by a mutant strategy D_1 , which prescribes to Send when in the role of truster, and to Breach when in the role of trustee.

The incumbent strategy C yields an expected payoff equal to

$$u[C, \varepsilon D_1 + (1 - \varepsilon)C] = \frac{1}{3} (1 - \varepsilon^2)36 + \frac{2}{3} 30 = 32 - 12\varepsilon^2$$

while the “mutant” strategy D_1 yields an expected payoff equal to

$$u[D_1, \varepsilon D_1 + (1 - \varepsilon)C] = \frac{1}{3} (1 - \varepsilon^2)36 + \frac{2}{3} [\varepsilon 48 + (1 - \varepsilon)30] = 32 - 12\varepsilon^2 + 12\varepsilon.$$

Hence,

$$u[C, \varepsilon D_1 + (1 - \varepsilon)C] < u[D_1, \varepsilon D_1 + (1 - \varepsilon)C], \quad \forall \varepsilon > 0.$$

Along the same lines, one could show that none of the four possible pure strategies available in this game is evolutionary stable.²⁵

A.2. Coordination variation

A similar reasoning shows that the cooperative strategy C is evolutionary stable, as it cannot be invaded neither by any of the three other possible pure strategies, nor by the mixed strategy S which prescribes to Send when in the role of a truster, and to Return with probability $\frac{1}{3}$ when in the role of a trustee.

The intuition is that any mutant strategy which prescribes to play Keep when in the role of a truster grants the truster a payoff of 20 units, which – for ε low enough – is lower than the expected payoff the truster gets by playing Send (which is at least equal to $36 - 36\varepsilon^2$ units). On the other hand, any strategy that prescribes to Breach when in the role of a trustee yields the trustee an expected payoff of $48\varepsilon + 20(1 - \varepsilon)$ when the truster chooses to Send, which – for ε low enough – is lower than the sure payoff of 30 the trustee would get by playing the cooperative strategy C .

Appendix B. Dictator games and other-regarding preferences

In each session, participants faced three 3-player modified dictator games (DGs). In each modified dictator game, each participant (red player) had to choose how to allocate tokens among himself and the other two players in the group (black and white players). In each game – DG1, DG2, and DG3 – the dictator faced six alternative allocations: 1, 2, ..., 6. The strategy method was used and each participant played as if he was the dictator (red player) in all three games. At the end of the session, one game was selected at random for payment and roles were assigned.

In DG1 the dictator allocated MUs between herself and two others, and chose an option that gradually ranged between E which grants 160 MUs to the dictator and 160 MUs to each of the other two persons and W (160 MUs, 130 MUs, 340 MUs). The dictator's earnings were always equal to 160 tokens and a merely self-interested dictator would be indifferent among all the available allocations. While allocation E ensured equal earnings to all three group members, allocation W delivered the highest sum of earnings for the group (W). The choice of the dictator in DG1 – i.e. option 1, 2, ..., 6 – revealed his relative preferences for equality over group wealth.

In DG2 the allocations ranged between S (190 MUs, 40 MUs, 250 MUs) and W. The dictator faced a tradeoff between self-interest and group wealth. Notice that allocation W was identical in both DG1 and DG2. Moreover, group wealth in each allocation was identical across DG1 and DG2. In DG2, however, dictator's earnings vary from 160 (allocation W) to 190

²⁵ The exact computations are available from the authors, upon request.

Table B1

Classification of subjects into types, according to their concerns for group wealth and equality.

<i>Treatment</i>	Baseline active	Coordination active	Baseline passive	Coordination passive	Total
Strong concerns for group wealth	25	21	11	10	67
Strong concerns for equality	17	22	12	7	58
Others	18	14	7	10	49
Total	60	57	30	27	174

(allocation S) tokens.²⁶ A self-interested dictator would always choose allocation S over all other allocations in DG2. In contrast, in DG3 the choice ranged between S and E. Group wealth was kept constant and the dictator faced a tradeoff between self-interest and equality. A self-interested dictator will always choose allocation S and earn 190 tokens, while a dictator concerned with inequality may choose allocation E and earn 160 tokens.

Table B1 classifies types according to their concerns for group wealth and equality. We say that a participant displays a strong concern for group wealth when she chooses options leaning toward W (4, 5, or 6) in both DG1 and DG2. Symmetrically, a participant expresses a strong concern for equality when she chooses options leaning toward E (1, 2 or 3) in DG1 and leaning toward E (4, 5 or 6) in DG3. No statistically significant difference emerges in the distribution of types across treatments (a chi-squared test returns a *p*-value of 0.756).

Appendix C. Beliefs and risk attitudes

Is there a correlation between trust and pre-play beliefs about others' trustworthiness? That is, do participants correctly anticipate the role of derived trustworthiness? To assess whether trusters correctly anticipate that trustees return more when penalties for mis-coordination are in place, we look at beliefs about others' trustworthiness as reported by our participants before playing the Collective Trust Game. For simplicity, in the present discussion, we limit our analysis to beliefs about trustworthiness in *Active* treatments (beliefs, hereafter).

Result A1. Ex-ante beliefs on trustworthiness are not significantly different across treatments.

Whereas observed levels of trustworthiness are significantly different across treatments, pre-play beliefs are not: on an average, participants expected 39% of the trustees to play Return in the *Baseline-Active* and 44% in the *Coordination-Active* treatment ($p=0.278$, two-sample Wilcoxon rank-sum test, $N_1 = 60$ and $N_2 = 57$, two-sided). The evidence suggests that participants did not correctly anticipate behavioral differences between the two treatments.²⁷

One can also compare beliefs and actions at the individual level. As one would have expected, we find that the higher the pre-play belief about others' trustworthiness, the higher the initial level of trust (see Fig. C1). To study the coherence between pre-play beliefs and initial trust behavior, we restrict our analysis to the first choice a participant made when playing as truster. In *Coordination-Active*, participants who first chose Send expect a larger fraction of the trustees to Reciprocate as compared to participants who first chose Keep and the difference is statistically significant according to a two-tailed Wilcoxon rank-sum test ($p=0.019$, $N_1 = 39$ and $N_2 = 18$); the same pattern, albeit not significant, was observed also in *Baseline-Active* ($p=0.197$, $N = 30$ and $N = 30$). Pre-play beliefs are also strongly correlated with initial trustworthy choices; while in *Coordination-Active* this positive correlation can be easily interpreted – there is a monetary cost associated to mis-coordination – the same is not true for *Baseline-Active*. In this case, the difference is highly significant for both *Coordination-Active* (two-tailed Wilcoxon rank-sum test, $p < 0.001$, $N = 32$ and $N = 25$) and *Baseline-Active* (two-tailed Wilcoxon rank-sum test, $p < 0.001$, $N_1 = 37$ and $N_2 = 23$).

Is there a correlation between trust and risk attitudes? We found that beliefs play an important role in shaping initial behavior, but other forces, such as risk preferences, may influence trusters' choices, as well.

Result A2. Risk attitudes significantly correlate with trust in the *Baseline* and not in the *Coordination* treatment.

To shed further light on the determinants of trust, we ran a regression on trusters' choices against their risk attitudes and beliefs. Table C1 presents probit estimations separately for *Coordination-Active* and *Baseline-Active*. Two dummies for risk attitudes and for beliefs are included in all specifications. The dummy *Strongly risk averse* (*Risk neutral or loving*) takes value

²⁶ Allocations W through S in DG2 were designed to have the same level of inequality as measured according to Fehr and Schmidt (1999)'s model, under the assumption of equal weights for disadvantageous and advantageous inequality. The difference in inequality measured according to Bolton and Ockenfels (2000)'s model is also minimal in these two allocations.

²⁷ Sapienza et al. (2007) maintains that beliefs are participant to updating as the game progresses due to learning; in line with this conjecture, our data show a marked difference between beliefs elicited in the first and the last period in the *Coordination-Active* treatment. Before the final period, the difference in beliefs between treatments is substantial and significant: on an average participants expected 34% of the trustees to play Return in the *Baseline-Active* and 58% in the *Coordination-Active* treatment ($p=0.043$, two-sample Wilcoxon rank-sum test, $N_1 = N = 4$, two-sided). The stability of beliefs over time in the *Baseline-Active* treatment should not come as a surprise since beliefs were already accurate in the first period. It is worth noticing that beliefs in the last period and actual behavior were remarkably close in both treatments; that is, participants correctly learned norms and conventions in use.

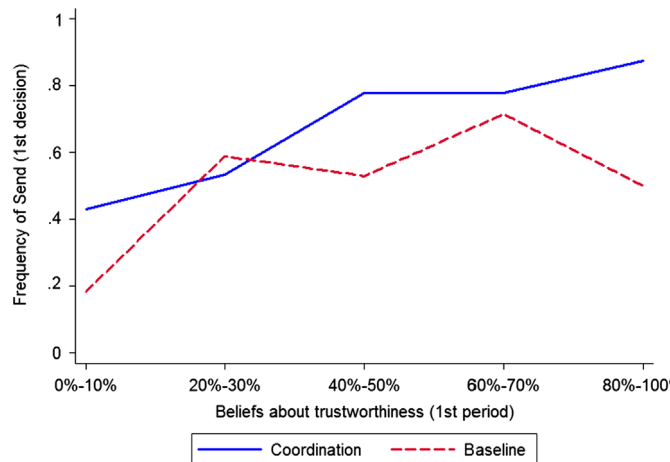


Fig. C1. Initial actions and pre-play beliefs: active treatments.

Table C1

Beliefs and risk aversion on trust: Active treatments.

Dependent variable: Send (1) Keep (0)	Coordination		Baseline	
	Model 1	Model 2	Model 3	Model 4
1/Period	−0.238*** (0.067)	−0.235*** (0.065)	−0.052 (0.113)	−0.053 (0.106)
Strongly risk averse	−0.042 (0.063)	−0.048 (0.067)	−0.143 (0.120)	−0.191* (0.114)
Risk neutral or loving	0.002 (0.069)	−0.048 (0.102)	0.376*** (0.114)	0.309** (0.140)
Low belief ($p=1$)	−0.080 (0.078)	−0.063 (0.072)	−0.209 (0.127)	−0.124 (0.136)
High belief ($p=1$)	0.026 (0.058)	−0.004 (0.070)	0.116 (0.141)	0.172 (0.130)
Demographics	No	Yes	No	Yes
N.obs.	570	570	600	600
Log likelihood	−231.713	−227.342	−315.084	−311.757

Notes: Marginal effects from probit regression with individual random effect. Active treatments only. For Demographics see notes to Table 4.

* indicates the significance at 10%.

** indicates the significance at 5%.

*** indicates the significance at 1%.

1 for participants whose choices in the lottery task is compatible with a relative risk aversion coefficient smaller than 0.50 (larger than 1.16), and 0 otherwise.²⁸ Similarly, we created two dummies for pre-play beliefs on others' trustworthiness; *Low belief* (*High belief*) is equal to 1 if a participant expects less than 1/3 (more than 2/3) of the trustees to be trustworthy.

In *Baseline-Active*, unlike in *Coordination-Active*, higher levels of trust are associated with more risk tolerance, and the correlation is statistically significant (Table C1). Marginal effects of pre-play beliefs are larger in *Baseline-Active* than in *Coordination-Active* treatments, where they can explain a smaller proportion of the variation of trust across participants. However, pre-play beliefs are not statistically significant neither in *Coordination* nor in *Baseline* treatments.

Appendix D. Instructions

In this Appendix we report the English translation of the original Italian instructions. In italics and squared brackets we report the parts included in the experimenter instructions, but not in the versions for the subjects.

Treatments : Baseline–Active and [Baseline–Passive]

Welcome! This study is part of a research project of the University of Bologna and it is financed by the European Commission.


²⁸ Average choice in the risky task was 3.00 (3.18) in *Baseline-Active* (*Coordination-Active*). A two-tailed Wilcoxon rank-sum reveals no differences in risk attitudes between the two treatments ($p=0.513$, $N=60$ and $N=57$).

You will earn money depending on your choices and on the choices of the other participants. You will be paid in private at the end of today's study.

Turn off your mobile phone. From this moment on, no form of communication among participants is allowed. In case you have a question, please raise your hand and one of us will come to your desk to answer it.

Follow the instructions carefully. In this study there are three parts; I am about to read instructions for Part 1.

Premi il bottone corrispondente all'opzione che preferisci.



opzione	se la pallina estratta è ARANCIONE guadagni:	se la pallina estratta è BIANCA guadagni:	contenuto dell'URNA
I	17.5	17.5	<div style="display: flex; justify-content: space-around;"> ●●●●●● </div> <div style="display: flex; justify-content: space-around;"> ○○○○○○ </div>
II	22.5	15.0	<div style="display: flex; justify-content: space-around;"> ●●●●●● </div> <div style="display: flex; justify-content: space-around;"> ○○○○○○ </div>
III	27.5	12.5	<div style="display: flex; justify-content: space-around;"> ●●●●●● </div> <div style="display: flex; justify-content: space-around;"> ○○○○○○ </div>
IV	32.5	10.0	<div style="display: flex; justify-content: space-around;"> ●●●●●● </div> <div style="display: flex; justify-content: space-around;"> ○○○○○○ </div>
V	37.5	7.5	<div style="display: flex; justify-content: space-around;"> ●●●●●● </div> <div style="display: flex; justify-content: space-around;"> ○○○○○○ </div>
VI	44.0	1.0	<div style="display: flex; justify-content: space-around;"> ●●●●●● </div> <div style="display: flex; justify-content: space-around;"> ○○○○○○ </div>

D.1. Instructions: Part 1

In this part, you have to choose among six different earnings options. Each option can produce either a high or a low earning. Look at the screen; for each option:

- the high earning is in the second column;
- the low earning is in the third column.

The high earning has a 50% probability to be realized and the low earning has a 50% probability to be realized.

What is your task? You have to choose your favorite option. Look at the screen: there is a button – I to VI – for each row. In order to choose, you have to press the button next to your favourite option. Touch the screen only with your fingers; pencils could damage the screen.

How are your earnings computed?

- at the end of today's study, two participants in this room will be selected at random; only the selected participants will receive a payment for this part. The payment can be either high or low;
- there will be an urn containing ten balls – 5 orange and 5 white balls;
- if an orange ball is drawn from the urn, the selected participants will get the high payment for the selected option;
- if a white ball is drawn from the urn, the selected participants will get the low payment for the selected option.

Let us consider an example with no consequences for your final earnings. Press button V. You will earn Euros 37.5 if an orange ball is drawn from the urn, while you will earn Euros 7.5 if a white ball is drawn from the urn. [*Have you pressed the button?*] A box to CONFIRM or CHANGE your choice has appeared on the screen; please, press CHANGE. Now you can change your choice. [OK] Now press the button III; you will earn Euros 27.5 if an orange ball is drawn from the urn, while you will earn Euros 12.5 if a white ball is drawn from the urn. Press CONFIRM. [*The choice cannot be changed anymore*]. Everyone press CONFIRM. Is everything clear? If there are no questions, we can start with Part 1.

D.2. Instructions: Part 2

In the present and subsequent parts, your earnings are expressed in tokens: tokens will be converted in Euros at the rate of 1 Euro for 40 tokens.

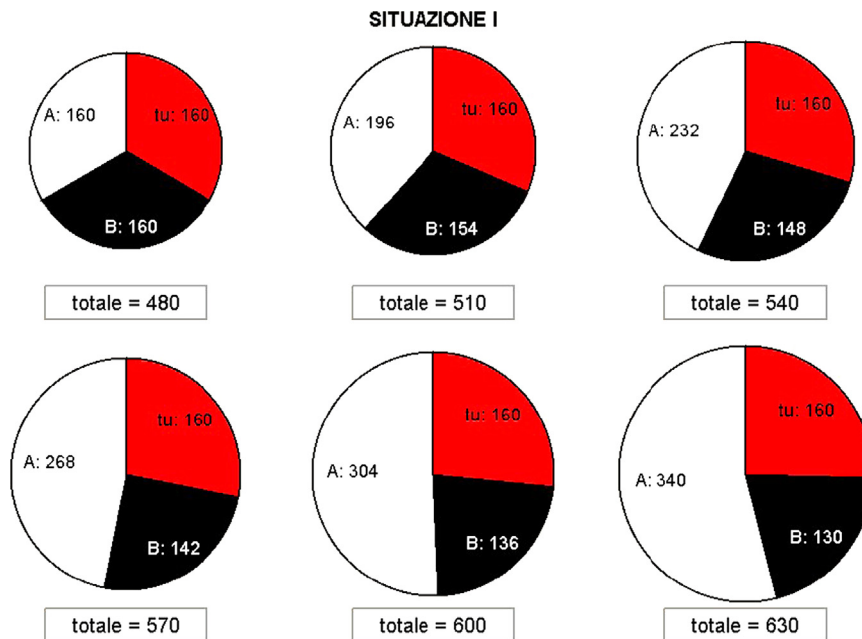
In this part, people in this room are randomly divided into groups of three; nobody can know the identity of the other members of the group.

Three situations will be presented in turn. What is a situation? Look at the screen, you can see an example of a situation. [Can you see six figures on the screen? OK] Each figure is divided in three slices; the red slice indicates your earnings, while the black and the white slices represent the earnings of the other people in your group.

For instance, in the top-left figure, your earnings, as well as the earnings of the other two members of your group, amount to 160 tokens. Let us consider another example; if you choose the bottom-right figure, you will earn 160 tokens, while one person in your group will earn 340 tokens and the other will earn 130 tokens. [Is everything clear?]

As you can see, different figures can have different dimensions. The sum of the earnings of each member of the group is displayed below each figure. [Is there any question?]

What is your task? You have to choose one of the six figures; in order to choose, you have to press the figure you prefer the most. As an example without consequences for your earnings, press the top-center figure. A box to “CONFIRM” or “CHANGE” your choice has appeared on the screen; the chosen figure is highlighted by a white box. Press CONFIRM. [Your choice cannot be changed now. Is there any question?]



How are your earnings computed? Every person in your group will make a choice for each situation. Among all the choices made within your group, only one randomly chosen choice will be implemented; the implemented choice can be your choice or the choice made by another member of your group.

What if your choice is chosen at random? Your choice will determine your earnings and the earnings of the other members of your group.

What if the choice of another person in your group is chosen at random? It can be the case that your earnings are different from the ones you chose. In this case, your earnings depend on the choice made by the selected person and by the color you have been assigned at random: either white or black. The choice will be randomly selected at the end of this part; therefore, you have to pay attention to all of your choices. [Is everything clear?]

Before starting, please answer a few questions.

D.3. Instructions: Part 3

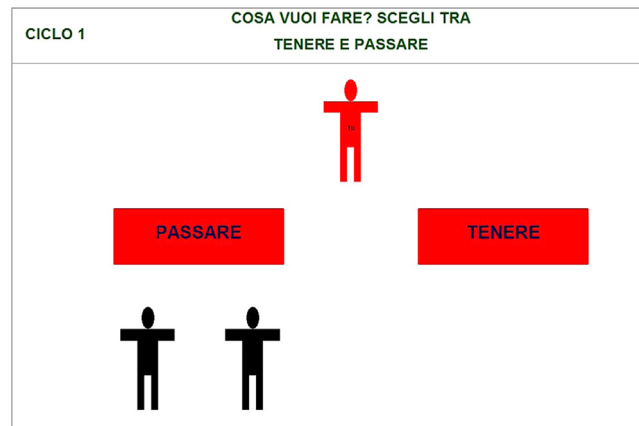
As in the previous part, people in this room are divided at random in groups of three people; nobody can know the identity of the other members of the group.

There are two different roles in each group: role A and role B. In each group, one person has the role A and two have the role B. Roles are randomly assigned by the computer.

What is the task for a role A person? Look at the screen, A has to decide between PASS and KEEP.

- If the person A decides to KEEP, everyone in the group earns 20 tokens. In this case, people playing as B do not have to take any decision;
- if the person A decides to PASS, she (he) earns either 0 or 36, depending on the choices made by people playing as B in the group.

As an example with no consequences for your earnings, please press the button PASS. You can now decide whether to CONFIRM or CHANGE your choice; please, press CONFIRM. [*The choice cannot be changed anymore. Everyone press CONFIRM.*]
 [The computer will draw a ball from an urn containing blue and yellow balls:



:

- if the ball is yellow, everyone in the group earns 20 tokens. In this case, people playing as B do not have to take any decision;
- if the ball is blue, A will earn either 0 or 36 depending on the choices made by people playing as B in the group.

The person playing as role A has no choice to make.

What is the task for B if A decides to PASS [if the ball is blue]? Look at the screen, the two participants with role B have to simultaneously choose between GIVE and KEEP. If A decides to PASS [if the ball is blue], how are the earnings computed?

- if both Bs decide to KEEP, Bs earn 48 tokens each and A earns 0 tokens;
- if both Bs decide to GIVE, Bs earn 30 tokens each and A earns 36 tokens;
- if one B decides to GIVE and the other decides to KEEP, Bs earn 30 tokens each and A earns 36 tokens.

Remember that everyone in the group earns 20 tokens if A decides to KEEP [if the ball is yellow].

Let us consider an example with no consequences for your earnings. In this example, you have been assigned to the role B. Press KEEP and then CONFIRM. [*The choice cannot be changed anymore. Everyone press CONFIRM.*]

In the following screen, you can see the final earnings for the group. In the present example, you have been assigned role B and you have decided to KEEP, while the other person who has been assigned to role B has decided to GIVE. Both you and the other person with the role B earn 30 tokens and A earns 36 tokens.

In this part there are 30 rounds with the same rules. In the upper-left part of the screen you can see the number of the current round. At the beginning of every round, new groups of three people are formed at random.

To sum up, in every round:

- a has to decide between PASS and KEEP;
- if A decides to KEEP [if the ball is yellow], the round ends and everyone earns 20 tokens;
- if A decides to PASS [if the ball is blue], the two people with role B have to choose between GIVE and KEEP and earnings are computed as explained above.

Earnings cumulate from round to round. *The composition of the urn has been decided in advance and does not depend on the choices people in this room will make.* [Is everything clear?] Before starting, please answer a few questions.

Appendix E. Supplementary data

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.euroecorev.2013.08.009>.

References

- Anderlini, L., Terlizzese, D., 2012. Equilibrium Trust. Technical Report.
 Bacharach, M., Gambetta, D., 2001. Trust in signs. *Trust in Society* 2, 148–184.

- Bacharach, M., Guerra, G., Zizzo, D., 2007. The self-fulfilling property of trust: an experimental study. *Theory and Decision* 63, 349–388.
- Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10, 122–142.
- Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. *American Economic Review* 90, 166–193.
- Burks, S., Carpenter, J., Verhoogen, E., 2003. Playing both roles in the trust game. *Journal of Economic Behavior and Organization* 51, 195–216.
- Camerer, C., 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.
- Cason, T., Mui, V., 2007. Communication and coordination in the laboratory collective resistance game. *Experimental Economics* 10, 251–267.
- Cassar, A., Rigdon, M., 2011. Trust and trustworthiness in networked exchange. *Games and Economic Behavior* 71, 282–303.
- Charness, G., Dufwenberg, M., 2006. Promises and partnership. *Econometrica* 74, 1579–1601.
- Coleman, J., 1990. *Foundations of Social Theory*. The Belknap Press of Harvard University Press, Cambridge, Massachusetts.
- Cox, J., 2004. How to identify trust and reciprocity. *Games and Economic Behavior* 46, 260–281.
- Crosen, R.T., 2000. Thinking like a game theorist: factors affecting the frequency of equilibrium play. *Journal of Economic Behavior and Organization* 41, 299–314.
- Dasgupta, P., 1988. Trust as a commodity. In: Gambetta, D. (Ed.), *Trust. Making and Breaking Cooperative Relations*, Basil Blackwell, New York, pp. 49–72.
- Eckel, C., Wilson, R., 2004. Is trust a risky decision? *Journal of Economic Behavior and Organization* 55, 447–465.
- Eckel, C.C., Grossman, P.J., 2002. Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior* 23, 281–295.
- Eckel, C.C., Grossman, P.J., 2008. Forecasting risk attitudes: an experimental study using actual and forecast gamble choices. *Journal of Economic Behavior and Organization* 68, 1–17.
- Fehr, E., 2009. On the economics and biology of trust. *Journal of the European Economic Association* 7, 235–266.
- Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114, 817–868.
- Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10, 171–178.
- Greiner, B., 2004. *The Online Recruitment System ORSEE 2.0—a Guide for the Organization of Experiments in Economics*. University of Cologne, Working Paper Series in Economics, No. 10.
- Hirschman, A.O., 1984. Against parsimony: three easy ways of complicating some categories of economic discourse. *American Economic Review Papers and Proceedings* 74, 88–96.
- Houser, D., Schunk, D., Winter, J., 2010a. Distinguishing trust from risk: an anatomy of the investment game. *Journal of Economic Behavior and Organization* 74, 72–81.
- Houser, D., Vetter, S., Winter, J.K., 2010b. Fairness and Cheating. Discussion Paper Series of SFB/TR 15 Governance and the Efficiency of Economic Systems, Free University of Berlin, Humboldt University of Berlin, University of Bonn, University of Mannheim, University of Munich.
- Johnson, N., Mislin, A., 2011. Trust games: a meta-analysis. *Journal of Economic Psychology* 32, 865–889.
- McCabe, K., Rigdon, M., Smith, V., 2003. Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization* 52, 267–275.
- McEvily, B., Weber, R., Bicchieri, C., Ho, V., 2006. Can groups be trusted? An experimental study of trust in collective entities. In: Reinhard Bachmann, A.Z. (Ed.), *Handbook of Trust Research*, Edward Elgar, Cheltenham, UK, pp. 52–67.
- Mittone, L., Ploner, M., 2009. Social Effects in a Multi-Agent Investment Game. An Experimental Analysis. CEEL Working Papers.
- Regner, T., Riener, G., 2011. Motivational Cherry Picking. Jena Economic Research Papers.
- Rousseau, D., Sitkin, S., Burt, R., Camerer, C., 1998. Not so different after all: a cross-discipline view of trust. *Academy of Management Review* 23, 393–404.
- Sapienza, P., Toldra, A., Zingales, L., 2007. Understanding Trust.
- Schecter, L., 2007. Traditional trust measurement and the risk confound: an experiment in rural Paraguay. *Journal of Economic Behavior and Organization* 62, 272–292.
- Schelling, T., 1960. *The Strategy of Conflict*. Harvard University Press.
- Sheremeta, R.M., Zhang, J., 2012. Three-Player Trust Game with Insider Communication. Working Papers, University of Zurich, Department of Economics.
- Weibull, J., 1997. *Evolutionary Game Theory*. The MIT Press.
- Weingast, B., 1997. The political foundations of democracy and the rule of law. *American Political Science Review*, 245–263.