European Economic Review 64 (2013) 129–146

Contents lists available at ScienceDirect

European Economic Review

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

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.

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).

0014-2921/$ - see front matter
http://dx.doi.org/10.1016/j.euroecorev.2013.08.009

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.