Games and Economic Behavior 102 (2017) 624–644

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Equilibrium trust ☆

Luca Anderlini [a,*], Daniele Terlizzese [b]

[a] Georgetown University, United States
[b] EIEF and Bank of Italy, Italy

ARTICLE INFO

Article history:
Received 5 November 2015
Available online 9 March 2017

JEL classification:
D80
D89
C79

Keywords:
Trust
Social norms
Multiple equilibria
Enforcement

ABSTRACT

Trusting beliefs can be exploited. A trustful player who is cheated too often, should start trusting less, until her beliefs are correct. For this reason we model trust as an *equilibrium* phenomenon. Receivers of an offer to transact choose whether or not to cheat. Cheating entails a cost, with an idiosyncratic component and a socially determined one, decreasing with the mass of players who cheat. The model either has a unique equilibrium level of trust (the proportion of transactions not cheated on), or two — one with high and one with low trust. Differences in trust can result from different fundamentals or from different equilibria being realized. Surprisingly, under certain conditions these two alternatives are partially identifiable from an empirical point of view. Our model can be reinterpreted with the cost of cheating arising from an enforcement mechanism that punishes cheaters in a targeted way using limited resources.

1. Introduction

1.1. Motivation and overview

Not many transactions are carried out using rotating security trays with money on one side and the purchased good on the other. Yet, without these devices or some equivalent arrangement, the standard selfish players that populate modern economic models would be unable to trade, barring repeated interaction or binding contracts. It is instead self-evident that trade among players flourishes way beyond what this hypothetical world would look like.

The lubricant that makes so many transactions take place is *trust*. This is the belief that economic players hold that the other side of the transaction will not behave in a completely opportunistic way, thus impeding mutually advantageous exchange.[1]

Our purpose here is to build a simple model of trust as an *equilibrium* phenomenon. The players' beliefs about not being cheated — their level of trust as we just defined it — should be endogenously determined in equilibrium, and hence *correct*.

☆ We are grateful to Helena Aten, Larry Blume, Martin Duwfenberg, Hülia Eraslan, Leonardo Felli, Dino Gerardi, Hugo Hopenhayn, Guido Kuersteiner, Roger Lagunoff, George Mailath, Facundo Piguillem, Aleh Tsyvinski, Bill Zame and seminar participants for helpful comments. We also thank three anonymous referees and the Advisory Editor of this Journal for constructive criticism. Luca Anderlini thanks EIEF for their generous hospitality. All errors remain our own.

* Corresponding author at: Department of Economics, Georgetown University, 37th and O Streets, Washington, DC 20007, USA.
E-mail address: luca@anderlini.net (L. Anderlini).

[1] We define trust as a belief, in line with Charness and Dufwenberg (2006), rather than adopting the behavioral definition advocated by Fehr (2009). Since in equilibrium there will be a one-to-one mapping between beliefs and behavior, the distinction for us is largely immaterial. For ease of exposition, we postpone until Subsection 1.2 a discussion of specific contributions from the large literature on the sources and effects of the presence of trust.

http://dx.doi.org/10.1016/j.geb.2017.02.018

We focus on equilibrium beliefs because of a basic tension between "trusting beliefs", and consequent "trusting behavior", and the incentives to cheat of other players in society. Trusting beliefs can be exploited. However, a trustful player should not be cheated too much; if she is, she would change her belief and start trusting less. This tension drives our interest for what an equilibrium model of trust can generate.

We are not the first to stress the need for an equilibrium analysis of trust. Within the framework of psychological games, Huang and Wu (1994) and Dufwenberg (1996, 2002) analyze equilibrium trust in models that bear many resemblances to ours. What is new in this paper is an analysis of both the extensive and intensive margins of trusting behavior (not only whether to trust but also how much) and elements of heterogeneity (not all agents deserve to be trusted equally).[2] The conjunction of these two features allows us to explore novel comparative statics which we think shed light on the role of multiple vs unique trust equilibria in the interpretation of empirical evidence.

To model trust as an equilibrium phenomenon we place a "cost of cheating" in the players' utility function. The cost of cheating has two components. One which is an exogenous, idiosyncratic characteristic of each player, and another which is socially determined by the behavior of others. The less common cheating behavior is in society, the higher is the cost of cheating for individual players. This feedback component of the cost of cheating is a central ingredient of our analysis. We interpret it as reflecting a social norm, determined by the average behavior of others. It can also be given a somewhat different psychological underpinning, in the context of the theory of psychological games (we postpone a discussion of this point to Subsection 1.2). The experimental evidence presented in Charness and Dufwenberg (2006) offers strong empirical support for the presence of a social feedback of a similar nature. Alternatively, we might think of the feedback as resulting from an enforcement technology, whose effectiveness depends, for given resources, on the average behavior. We return to this alternative interpretation in Section 6.

Our model is deliberately kept simple in the extreme. However, as we will argue below, the flavor of our results is independent of many of the stark features of our model. Two features of our set-up are worth mentioning at the outset.

First, our model is *not* dynamic, nor should it be interpreted as a "reduced form" of a dynamic set up. Repeated interactions can and have been used successfully to generate cooperative and trust-like behavior, together with a very large variety of other equilibria. However, both in real life and in laboratory experiments trust seems to emerge even in one-shot interactions, in situations that would seem to call for "swivel-tray trading" if trust were not present. To sharpen our understanding of these situations, we work with a model that avoids reputational issues and more generally repeated interaction altogether.

Secondly, our set up is a "one-sided" model of trust, with one player making a proposal, and a responder who can decide whether to cheat or not. Admittedly, trust is often required on both sides of a transaction, and in those cases trust should be modeled as a two-sided phenomenon in which both players in a match have a choice of whether to cooperate or not. Our approach is therefore restrictive. There is, however, a large and growing experimental literature on trust games and reciprocity that explores the emergence of trust in situations akin to the one we model.

In spite of its extreme simplicity, we believe our set up provides a rich enough framework to address the well documented diversity of levels of trust in different societies. Indeed, depending on the configuration of preferences and other parameters, our model either generates a unique equilibrium (with a single equilibrium level of trust), or two equilibria, one with a higher and another with a lower level of trust. Moreover, by varying the parameters of the model, higher or lower levels of equilibrium trust can be obtained without switching across different equilibria. Given this rich set of possible equilibrium outcomes, the model allows us to frame in a natural way what seems a key question concerning the levels of trust in different societies. When we observe different levels of trust across different societies, is this due to a difference in the *fundamental parameters* that underpin the different societies, or is it possible that the *same fundamental parameters* give rise to *different equilibria*? Obviously these two possibilities have vastly different policy implications, and it is therefore important to have a framework in which they can be made precise, and hence disentangled.

Of course, unless the multiplicity can be somehow validated empirically, models with multiple equilibria are no more than a theoretical benchmark providing an — albeit important — only *potential* interpretation of reality. A key insight from our model is that the two cases mentioned above, under some conditions, are *partially identifiable*: if different levels of trust result from multiple equilibria, then the level of trust must be negatively correlated with the size of individual transactions. A positive correlation can only emerge if different levels of trust were to result from differences in the parameters of the model.[3] To our knowledge, the possibility to empirically disentangle multiple vs. unique equilibrium regimes is new.

As we mentioned above, our model can be reinterpreted so that the social feedback component of the cost of cheating comes instead from an enforcement technology that punishes cheating using limited resources in a targeted way. We find that, in the multiple equilibria regime, an infinitesimal increase in the resources devoted to enforcement can yield a discontinuous increase in the level of total activity in the economy. In the single equilibrium regime, instead, the level of activity changes continuously with the resources spent in the enforcement technology.

[2] Heterogeneity is also considered by Attanasi et al. (2016). See footnote 8.

[3] Since the term "partial identification" has been used before Phillips (1989), it is useful to be precise as to the meaning we give it here. The identification is *partial* in the sense that we cannot rule out that two equilibria, corresponding to two sets of parameters, entail a negative correlation between trust levels and size of individual transactions. Therefore, while a positive correlation excludes that the different levels of trust results from multiple equilibria with unchanged parameters, the observation of a negative correlation is inconclusive.