European Economic Review 77 (2015) 102–116

Contents lists available at ScienceDirect

European Economic Review

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

Trust, trustworthiness and the consensus effect: An evolutionary approach

Fabrizio Adriani [a,*], Silvia Sonderegger [b]

[a] Department of Economics, University of Leicester, University Road, Leicester, LE1 7RH, UK
[b] University of Nottingham and CeDEx, UK

ARTICLE INFO

Article history:
Received 11 February 2014
Accepted 8 April 2015
Available online 2 May 2015

JEL classification:
A13
C73
D02
D03
D82
Z1

Keywords:
Endogenous preferences
Trust
Consensus effect
Deterrence
Retribution
Crowding out

ABSTRACT

People often form expectations about others using the lens of their own attitudes (the so-called *consensus effect*). We study the implications of this for trust and trustworthiness in an evolutionary model where social preferences are endogenous. Trustworthy individuals are more "optimistic" than opportunists and are accordingly less afraid to engage in market-based exchanges, where they may be vulnerable to cheating. Depending on the distribution of social preferences in the population, the material benefits from greater participation may compensate for the costs of being trustworthy. By providing an explicit account of how individuals form and revise their beliefs, we are able to show the existence of a polymorphic equilibrium where both trustworthiness and opportunism coexist in the population. We also analyze the effect of enforcement, distinguishing between its role as deterrence of future misbehavior and as retribution for past misbehavior. We show that enforcement aimed at deterring opportunistic behavior has ambiguous effects on social preferences. It may favor the spreading of trustworthiness (*crowding in*), but the opposite (*crowding out*) may also occur. By contrast, crowding out never occur when punishment is merely intended as retribution.

1. Introduction

People tend to think that others are like them. Nice guys tend to think that others are nice, while crooks believe that other people have similarly shifty personalities. This consensus effect has long been recognized by psychologists, at least since the seminal paper by Ross et al. (1977). Economists have also increasingly started to document and to pay attention to this phenomenon.[1]

The aim of this work is to study the implications of the consensus effect for the long-term evolution of preferences. We consider a setup where individuals either have preferences that only reflect their selfish material welfare (*Opportunists*) or have other-regarding/principled preferences (*Unselfish*). People are randomly matched to play a trust game in which trusting is optimal only when one's counterparty is unselfish. However, individual preferences are private information, and, thus, players decide to trust or not based on their (possibly heterogeneous) beliefs about the composition of the overall population. Endowed with this setup, we use an indirect evolutionary approach (see Güth and Yaari, 1992) to ask what distributions of preferences are likely to arise in the long run.

E-mail address: fa148@le.ac.uk (F. Adriani).

[1] Experimental studies by economists include Selten and Ockenfels (1998), Engelmann and Strobel (2000, 2012), Sapienza et al. (2010), Blanco et al. (2009, 2011), Gächter et al. (2012), Costa-Gomes et al. (2010) and Ellingsen et al. (2010).

http://dx.doi.org/10.1016/j.euroecorev.2015.04.003
0014-2921/

Previous work has already established that, when preferences are unobservable, the Unselfish type may be adaptive provided that beliefs about the population are type dependent. In particular, Orbell and Dawes (1991) suggest that if unselfish individuals have a higher propensity to believe that others are unselfish, they will be more inclined to interact with others. Depending on the actual composition of the population, higher participation propensity may afford an advantage to the Unselfish type which may compensate for the cost of foregoing lucrative opportunities for expropriating others (a cost which is borne by the Unselfish but not by the Opportunists). This implies that the Unselfish type is not necessarily outperformed by the Opportunists and can thus be evolutionarily successful. Gamba (2013) provides a related argument.

These accounts for the survival of other-regarding preferences – Orbell and Dawes's (1991), Gamba's (2013) – are appealing because they do not rely on preferences being observable. This stands in contrast to most of the literature on the indirect evolutionary approach.[2] On the other hand, these models predict that those with unselfish preferences will strictly outperform the Opportunists whenever their share of the population is above some critical value. Any payoff monotone dynamics would thus necessarily lead to a monomorphic population. As a result, these theories fail to account for the considerable heterogeneity in behavior extensively documented by the experimental literature. Indeed, as argued by Samuelson (2005)

> "Perhaps one of the most robust findings to emerge from experimental economics is that (..) heterogeneity is widespread and substantial. Despite this, heterogeneity has often not played a prominent role in many theoretical models."

Our paper advances the literature in two respects. First, the consensus effect is explicitly derived from rational belief formation based on introspection, as in Dawes (1989) and subsequent literature (Goeree and Großer, 2006; Vanberg, 2008). Second, we let players observe an external signal on the distribution of preferences in the population before playing. The combination of these two elements generates our key result, namely that a polymorphic population (where unselfish and opportunistic preferences coexist) may be stable. In our framework, heterogeneity emerges endogenously, as an equilibrium feature. Our theoretical analysis is thus one of the few to account for heterogeneity in behavior.[3]

In a nutshell, the key forces in our model can be described as follows. Consider an environment where, thanks to higher participation propensity, the Unselfish type is (initially) more successful than the Opportunistic type. As the proportion of unselfish individuals increases, the risk of being cheated is reduced. This effect increases the fitness of the Unselfish (who have a higher propensity to trust) more than that of the Opportunists. There are however countervailing forces that set a natural upper bound to the share of unselfish individuals. As the Unselfish type spreads, all players (including the Opportunists) become more likely to observe objective evidence suggesting that trusting is indeed optimal. The Opportunists become accordingly more willing to trust and, consequently, participation propensities become less type-dependent. At the same time, higher participation rates (of both types) increase the scope for cheating, thus boosting the Opportunists' fitness. In essence, the very prevalence of unselfish individuals undermines their evolutionary advantage. This results in a stable polymorphic population where unselfish individuals do materially as well as opportunistic ones.

Although quite intuitive, these effects can only be captured through an explicit account of how players form and revise their (type dependent) beliefs. In this paper, we focus on the true consensus effect, which is consistent with Bayesian learning and a common prior.[4] This is obtained by relaxing the standard assumption that the distribution of types within a population is known by the players. When the distribution of types is unknown, it becomes rational for individuals to use their own types to make inferences about the overall population. This is precisely what happens in our setup; the share of unselfish individuals in the population is not perfectly observed and, hence, the (Bayesian) beliefs about the composition of the overall population are type-dependent.[5]

The second contribution of our paper focuses on the interaction between ethical attitudes and institutions aimed at sanctioning/preventing opportunistic behavior. In particular, we consider the extreme cases of external punishment purely aimed at *deterring* opportunistic behavior and that of punishment intended as mere *retribution* for past misbehavior. We find that deterrence always increases welfare in the short run (i.e. keeping the distribution of types fixed), but has ambiguous long term effects (when the distribution of types is endogenous). In the long term, deterrence makes participation decisions more similar across types, thus crowding out other regarding preferences.[6] For some parameter values, this may lead to more cheating and lower welfare. Retribution has somewhat opposite effects: it entails welfare costs in the short run (since the welfare of cheaters is reduced), but generates a more desirable distribution of preferences in the long run.

[2] See e.g., Robson (1990). An exception to this is Huck and Oechssler (1999), who consider a setup where preferences are unobservable. However, in their setup players observe the composition of the population from which the opponent is drawn.

[3] Stable polymorphisms of both altruistic and selfish individuals may arise in models with local interactions. See Cohen and Eshel (1976) and Eshel et al. (1998). The mechanism at work in these models is very different from ours.

[4] While the importance of the consensus effect is well established, its interpretation is more controversial. Some psychologists claim that people systematically overestimate the extent to which others are similar to them – the so-called "false consensus effect". Others – such as Dawes (1989), Goeree and Großer (2006) or Vanberg (2008) – argue that this tendency is compatible with a common prior and Bayesian learning – hence, the terminology "true consensus effect".

[5] Type-dependent beliefs also feature in Ellingsen and Johanneson (2008). In Adriani and Sonderegger (2009) the consensus effect arises as an equilibrium feature of a game where parents select the values to instil in their children.

[6] See e.g. Frey (1997) and Bénabou and Tirole (2003) for theoretical analyses of motivation crowding out, and Frey and Jegen (2001) for a survey of empirical evidence. Huck (1998) and Bar-Gill and Fershtman (2004, 2005) build models where, as in ours, preferences are derived endogenously and may be "crowded out" in the long-run by the institutional environment. However, the mechanisms at work are very different from ours.