Trust, Introspection, and Market Participation: an Evolutionary Approach*

Fabrizio Adriani
SOAS - University of London

Silvia Sonderegger
University of Bristol, CMPO

July 8, 2009

Abstract

We build a model where *introspection* matters – i.e., people rationally form expectations about others using the lens of their own attitudes. Since trustworthy individuals are more "optimistic" about people than opportunists, they are less afraid to engage in market-based exchanges, where they may be vulnerable to opportunistic behavior. Within this context, we use an indirect evolutionary approach to endogenize preferences for trustworthiness. In some cases, the material rewards from greater market participation may outweigh the material disadvantages from foregoing lucrative expropriation opportunities. This implies that trustworthiness may be evolutionary stable in the long-term. Although stricter enforcement (that limits the scope for opportunistic behavior) does in some cases favor the spreading of preferences for trustworthy behavior (*crowding in*) we show that the opposite (*crowding out*) may also occur. Our findings are consistent with recent empirical evidence.

Jel Codes: C73, D02, D81, D82, Z1.

Keywords: Endogenous Preferences, Trust, Introspection, Institutions, Enforcement, Crowding Out.

*We thank Simon Burgess, Ellen Greaves, Francesco Giovannoni, Luca Deidda, Luigi Guiso, Steffen Huck, Fabrizio Mattesini, Larry Samuelson, Robert Waldmann and seminar participants at University of Aberdeen, University of Bristol, University of Rome "Tor Vergata", and SOAS, University of London for comments. We owe special thanks to Ken Binmore for discussions and encouragement.

1 Introduction

Modern life often requires us to engage with strangers, who may potentially behave opportunistically. Within this context, the willingness to trust others becomes a pre-requisite for interaction. The importance of trust for economic exchange is documented by a recent paper by Guiso, Sapienza and Zingales (2008), who show that more trusting individuals are significantly more likely to invest in the stock market, even after accounting for risk aversion and stock-market optimism. This suggests that an individual's willingness to trust others has important consequences for his economic well-being.

The issue of trust is intimately related with that of intrinsic motivation, or ethical attitudes within society. When "honest" ethical attitudes are widespread, the risk of being expropriated is low, even in the absence of explicit enforcement measures and/or reputation concerns. Hence, trusting others is optimal. By contrast, when honest ethical attitudes are rare, the risk of being expropriated is high, and trusting others is no longer optimal.

This paper is concerned with a number of specific questions that arise within this context. What determines trusting behavior in individuals? What ethical attitudes are likely to emerge over time? What is the relationship between ethical values and external incentives? Although these issues have traditionally been shunned by theoretical economists, in recent years a growing theoretical literature has emerged. Examples include Huck (1998), Bisin and Verdier (2001), Bohnet, Frey and Huck (2001), Hauk and Saez-Martí (2002), Francois and Zabojnik (2005), Corneo and Jeanne (2009), Francois (2008), Tabellini (2008). Our paper adds to this literature, by providing novel insights into these questions. While the questions we wish to address are to some extent applied, our approach is closely related to the game-theoretic literature on the evolution of preferences, such as Binmore (1994 and 2005), Robson (2001), Samuelson (2004), Samuelson and Swinkels (2006), Rayo and Becker (2007).

A novel feature of our model is that *introspection* matters – that is, people form expectations about others using the lens of their own attitudes.[1] Psychologists have long recognized that there is a systematic relationship between people's own characteristics

[1]Introspection also plays a crucial role in Adriani and Sonderegger (2009). However, the focus of that work is on parents' incentives to instill pro-social values in their children.

and their beliefs about the characteristics of others. Starting from the seminal paper by Ross, Greene and House (1977), a vast psychology literature has emerged on this subject. Economists have also recently started to pay attention introspection, especially in relation with trust. A recent experiment by Sapienza, Toldra and Zingales (2007) suggests that people playing a trust game tend to extrapolate their opponent's behavior from their own.

While the importance of introspection is well established, the implications of this tendency are more debated. Some psychologists claim that it may induce people to systematically overestimate the extent to which others are similar to them – indeed, they refer to it as the "false consensus effect". In this paper, we will follow Dawes (1989), and restrict attention to the *rational* component of introspection.[2]

A detailed description of our model can be found in section 2. We assume that, although it is common knowledge that an individual may either be opportunistic or trustworthy, the precise share of each type in the population is unknown. A Bayesian individual will therefore look at the way *she* would behave in a certain situation in order to make predictions about the way *her counterparty* is likely to behave in the same situation (i.e., she will use introspection). This implies that individuals are more likely to engage in market interactions when they are themselves trustworthy. Hence, introspection generates a *selection effect*, since individuals become more or less likely to participate in market-based exchanges depending on their ethical attitudes. Butler, Giuliano, and Guiso (2009) provide experimental evidence indicating that this selection effect is sizeable. Moreover, it persists even when individuals are exposed to (imperfect) information about the pool of players from which their opponent is drawn.[3]

A second key feature of our framework is that we let the long-term ethical attitudes in the population be determined endogenously, as the product of an evolutionary selection process. In section 3, we characterize the conditions under which preferences for trustworthy behavior are evolutionary stable and may therefore persist in the long-run. This may appear surprising at first glance, since trustworthy individuals fail to expropriate others, even when they could get away with it. However, the selection effect resulting

[2]See also Vanberg (2008).

[3]Using data from the European Social Survey, the same authors also show that trust (measured on a 0-10 scale by the survey's questionnaire) is highly heterogeneous, even within the same country. In section 4, we informally discuss how a Bayesian framework may generate asymptotic differences in individual beliefs. See also Acemoglu, Chernozhukov, and Yildiz (2008) for a formal analysis.

from introspection affords a potential advantage to the trustworthy, since they are more likely to engage in market interactions.[4] We show that this may in some cases outweigh the material disadvantages from foregoing lucrative expropriation opportunities. Hence, it is possible that optimistic, trustworthy individuals may on average do materially better than pessimistic, opportunistic types. Butler, Giuliano and Guiso (2009) provide empirical evidence that supports this hypothesis. Their results show that trustworthy individuals tend to be more trusting and are therefore cheated on more often. However, these individuals also take fuller advantage of profitable trade opportunities. Overall, the evidence indicates that individuals who do materially better in life also exhibit a positive degree of trust/trustworthiness.

Since trusting behavior pays off only when trustworthy attitudes are sufficiently widespread, the selection effect favors trustworthy types only when their share in society is sufficiently large. Hence, our analysis shows that, although the trustworthy may spread, this may happen only once they have reached a critical mass. The model may therefore generate multiple evolutionary stable states.

An important insight of our analysis concerns the interaction between ethical attitudes and external enforcement aimed at limiting the scope of opportunistic behavior. The short-run effect on behavior of introducing external enforcement is analyzed in section 2.3. In section 3.3, we show that, although strong external enforcement does in some cases favor the spreading of trustworthy ethical attitudes in the long-run, this is not always the case. Strong enforcement may "crowd out" trustworthy ethical attitudes. As will become clear, this happens because strong enforcement weakens the selection effect. Hence, our results indicate that the short-run and long-run effects of different institutional environments may conflict with each other. While in the short run the distribution of preferences (ethical attitudes) is fixed, in the longer-term these evolve endogenously, and are therefore affected by the surrounding institutional environment. We provide an example of how institutional arrangements that are "good" in the short-term may actually turn out to be "bad" once the endogeneity of preferences is taken into account. Bohnet, Frey and Huck (2001) present experimental evidence for crowding out.

[4] Orbell and Dawes (1991) first noticed that pro-social individuals had a potential advantage in the fact that they had more optimistic beliefs and were thus more willing to engage in potentially beneficial interactions. However they did not consider the evolutionary implications of this advantage. In other words, the fraction of pro-social individuals is exogenous in their model, while we determine it endogenously.

Their paper is further discussed in section 5, which provides concluding remarks.

The interpretation of our key assumptions, and the robustness of our results to relaxing these assumptions are discussed in section 4.

2 Introspection and beliefs

The starting point of our analysis is that individuals are willing to take part in market exchanges only if they believe their counterparty to be trustworthy with a sufficiently high probability. However, in an anonymous market, access to direct individual-level information about the trustworthiness of the counterparty may be limited. Introspection – i.e., looking at the way *you* would behave in a certain situation in order to make predictions about the way *others* are likely to behave in the same situation – may therefore be a useful source of information.

In what follows, we present a model where beliefs – and introspection – emerge from standard Bayesian updating. The fact that introspection is important for shaping individual beliefs is widely acknowledged (see e.g. Singer and Fehr, 2005 and the papers mentioned in the introduction). Indeed, psychologists even argue that we systematically tend to give excessive weight to ourselves when making predictions (the so-called false consensus effect). Building a model that allows for this may prove impractical, though. If individuals suffer from a systematic bias when evaluating information, then they may also suffer from other types of biases or departures from rational decision-making. For this reason, we present a model that conforms to standard economic modelling.

2.1 Benchmark model

Principals We consider a trust game where a risk neutral individual (the *principal*) must decide whether to participate in an exchange with another individual (the *agent*) who may engage in opportunistic behavior. To fix ideas, suppose that the principal is a buyer and the agent is the seller. The agent can behave opportunistically by delivering a damaged good or by not delivering at all. If the principal chooses not to participate, she will save her money, which gives her a material welfare equal to $\alpha > 0$. If the principal chooses to trust and the agent behaves honestly, the principal will obtain $\theta > \alpha$. In contrast, if the agent behaves opportunistically, the principal obtains zero. Hence, in this latter

case, the principal would have been better off not participating in the exchange at all. We assume away all issues of reputation and concentrate on the case in which the agent is a complete stranger, randomly drawn from the population, and the principal-agent interaction is one-shot.

Agents If the principal chooses not to trust, the agent receives a material payoff equal to zero. In contrast, if the principal trusts the agent and chooses to participate, the agent obtains a material payoff of $\rho > 0$ if he behaves opportunistically. An agent behaving honestly receives zero. In the buyer/seller example, a seller behaving honestly would make no extra profit, while a seller who, say, refuses to deliver the good, would make a profit equal to ρ. Note that an individual can benefit from being trusted only if he behaves opportunistically. By making life as difficult as possible for agents who behave honestly, this assumption works against the result we want to prove. It is however a good assumption for expositional purposes, since it allows to abstract from direct rewards from honest behavior (e.g. reputation, reciprocity, etc.).

We also assume $\rho < \theta$ – i.e. engaging in opportunistic behavior is inefficient. In the buyer/seller example, the buyer may derive higher material welfare from consumption of the good than the seller, so that more surplus is generated if the good ends up in the buyer's hands rather than in those of the seller. As will become clear, this assumption is necessary for the long-term survival of preferences for trustworthy behavior.

An agent's material welfare is thus maximized by behaving opportunistically whenever possible. On the other hand, opportunistic behavior may entail a psychological cost for some individuals. More specifically, we assume that all individuals belong to one of two types: opportunistic (O) and trustworthy (T). Type O individuals only care about material welfare. In contrast with type O, type T individuals suffer a psychological cost when behaving opportunistically. We assume that this cost is sufficiently high to ensure that type T always behave honestly.[5] It is important to stress that although opportunistic behavior may only be undertaken by individuals acting as agents, trustworthiness or opportunism characterize all individuals (including those acting as principals). Moreover, since opportunistic agents cheat whenever they can, they are materially better off than

[5]This psychological cost may be direct– as a result of the internalization of a "honesty" norm– or may arise indirectly– e.g., because people may have a preference for keeping their word, as in Ellingsen and Johannesson (2004).

trustworthy agents, who never cheat. We refer to this as the opportunists' *expropriation advantage*.

Information We assume that all individuals are drawn from the same (infinitely large) population, and that this is common knowledge. A principal is therefore aware that the population (from which her agent is randomly drawn) contains both type T and type O individuals. However, the precise share of type T is not known with certainty. This is a crucial assumption of our model since it ensures a role for introspection. By looking at her own type, a principal can gather useful information about the likelihood that her agent is trustworthy.

We denote with π the share of type T in the population (so that $1-\pi$ is the share of type O). The principal's prior over π has a non-degenerate cumulative distribution $F(\pi)$ and a density $f(\pi)$ with support $\mathcal{P} \subseteq [0,1]$. In addition to the prior, the principal has two pieces of relevant information. First, she observes a noisy signal $x \in X$ about π. The signal x is meant to capture the information that the principal is able to collect about the composition of the society. This typically reflects past personal experiences and the observed behavior of individuals in one's social network. Conditional on π, x has density $g(x|\pi)$ and cumulative $G(x|\pi)$. We denote with $\mu(x)$ the expected value of π given the prior F and a realization x, and with $\sigma^2(x)$ the conditional variance.[6] We assume that $\mu(.)$ is increasing and that $\sigma^2(x) > 0$ for all $x \in X$. The role of the first assumption is straightforward, the role of the second is to ensure that no realization of x can perfectly reveal the true value of π.

In addition to the signal x, the second piece of information available to a principal is her own type, τ. Since the principal does not perfectly observe the value of π, her type can be used to make inferences about the agent's type. If the principal knew the share π of type T in the population, then she would expect a randomly drawn agent to be of type T with probability π, independently of her type. However, since π is unobservable, the expectations of a type T principal differ from those of a type O principal, as shown

[6] From Bayes' rule,

$$\mu(x) = \int_{\pi \in \mathcal{P}} \pi \frac{g(x|\pi)dF(\pi)}{\int_{u \in \mathcal{P}} g(x|u)dF(u)} \tag{1}$$

and

$$\sigma^2(x) = \int_{\pi \in \mathcal{P}} (\pi - \mu(x))^2 \frac{g(x|\pi)dF(\pi)}{\int_{u \in \mathcal{P}} g(x|u)dF(u)}. \tag{2}$$

below.

2.2 The relationship between trustworthiness and trust, and the selection effect of introspection

Denote with $h(\pi|x,\tau)$ the posterior distribution of π given *both* x and the principal's type $\tau_P = O, L$. For a type T principal

$$h(\pi|x,\tau_P = T) = \pi \frac{g(x|\pi)f(\pi)}{\int_{u\in\mathcal{P}} ug(x|u)dF(u)} = \frac{\pi\tilde{g}(\pi|x)}{\mu(x)} \tag{3}$$

where $\tilde{g}(\pi|x) = g(x|\pi)f(\pi)/\int_{u\in\mathcal{P}} g(x|u)dF(u)$ is the posterior when observing x but not τ_P. Similarly, for a type O principal

$$h(\pi|x,\tau_P = O) = (1-\pi)\frac{g(x|\pi)f(\pi)}{\int_{u\in\mathcal{P}}(1-u)g(x|u)dF(u)} = \frac{(1-\pi)\tilde{g}(\pi|x)}{1-\mu(x)} \tag{4}$$

The last two expressions show that the principal's beliefs about π depend on her own type. Denoting with $\tilde{G}$ the cumulative distribution associated with $\tilde{g}$, and with τ_A the agent's type, a type T principal believes that the agent is a type T with probability

$$\Pr(\tau_A = T|x,\tau_P = T) = \frac{\int_{\pi\in\mathcal{P}} \pi^2 d\tilde{G}(\pi|x)}{\mu(x)} = \mu(x) + \frac{\sigma^2(x)}{\mu(x)} \tag{5}$$

The same probability for a type O principal is

$$\Pr(\tau_A = T|x,\tau_P = O) = \frac{\int_{\pi\in\mathcal{P}} \pi(1-\pi) d\tilde{G}(\pi|x)}{1-\mu(x)} = \mu(x) - \frac{\sigma^2(x)}{1-\mu(x)} \tag{6}$$

Given a value of x, the principal believes the agent to be trustworthy with higher probability when she is herself a trustworthy type. Individuals thus project their own characteristics onto others.

Given (5) and (6), the expected net payoff U from participating for a type T principal is

$$E(U|x,\tau_P = T) = \left(\mu(x) + \frac{\sigma^2(x)}{\mu(x)}\right)\theta - \alpha \tag{7}$$

The equivalent for a type O is

$$E(U|x,\tau_P = O) = \left(\mu(x) - \frac{\sigma^2(x)}{1-\mu(x)}\right)\theta - \alpha \tag{8}$$

The difference in expected net payoffs between a type T and a type O principals can then be written as

$$E(U|x, \tau_P = T) - E(U|x, \tau_P = O) = \theta \frac{\sigma^2(x)}{\mu(x)(1-\mu(x))} > 0 \tag{9}$$

which is always positive. Hence, keeping everything else equal, trustworthy individuals are always (weakly) more willing to take part in market exchanges than opportunists. We call this the *selection effect* of introspection. Proposition 1 summarizes the results so far.

Proposition 1. *(Selection effect) Given the same realization of the signal x: (i) a trustworthy principal believes the agent to be trustworthy with a higher probability than an opportunistic principal (i.e., she "trusts" more) and (ii) a trustworthy principal expects a higher material (net) payoff from participating than an opportunistic principal.*

Proposition 1 shows that, provided $\sigma^2(x) > 0$, different types of individuals hold different beliefs about the trustworthiness of others. The differences in the two types' beliefs is proportional to $\sigma^2(x)/\left[\mu(x)(1-\mu(x))\right]$. To interpret this ratio, note that the numerator is a measure of the accuracy of the signal x. The denominator is a measure of the accuracy of the other signal available to an individual, namely her type τ.[7] Overall, a small value for the ratio indicates that introspection is a relatively poor signal of π. Hence, the beliefs about the agent's trustworthiness held by a principal of type T should *not* be much more optimistic than those of a type O. By contrast, a large value for the ratio suggests that introspection is a relatively accurate signal. As a result, the beliefs of a principal of type T should be *much more* optimistic than those of a type O.

Our model thus predicts that differences in trusting attitudes may translate into different attitudes towards market participation. Butler, Giuliano, and Guiso (2009) and Sapienza, Toldra, and Zingales, (2007) provide experimental evidence indicating that

[7]Let the random variable $\boldsymbol{\tau}$ be equal to 1 if $\tau = T$ and zero otherwise, so that $E(\boldsymbol{\tau}|\pi) = \pi$ and $Var(\boldsymbol{\tau}|\pi) = \pi(1-\pi)$. Applying the law of total variance,

$$Var(\boldsymbol{\tau} \mid x) = E\left(Var(\boldsymbol{\tau} \mid \pi) \mid x\right) + Var(E(\boldsymbol{\tau} \mid \pi) \mid x) \tag{10}$$

Straightforward calculations show that the first term in (10) can be written as

$$E\left(Var(\boldsymbol{\tau} \mid \pi) \mid x\right) = \mu(x)(1-\mu(x)) - \sigma^2(x). \tag{11}$$

Moreover, since $E(\boldsymbol{\tau} \mid \pi) = \pi$, the second term in (10) is equal to $\sigma^2(x)$. Hence, $\mu(x)[1-\mu(x)]$ is equal to $Var(\boldsymbol{\tau} \mid x)$.

trustworthy individuals are significantly more trusting. The first of these studies also analyzes the relationship between trust and economic performance, using data at the individual level. The evidence suggests that more trustworthy individuals tend to be cheated more often. On the other hand, less trustworthy individuals – who are accordingly less trusting – tend to miss too many profit opportunities.[8]

2.3 Enforcement

We now extend the benchmark case to analyze the role of institutions. The institutional environment determines the extent to which market interactions can rely on external enforcement (enforcement in short). Enforcement acts as an exogenous limitation to the agent's ability to cheat the principal: an agent behaving opportunistically is able to expropriate the principal only with probability $1 - \phi$. More precisely, suppose that the specific circumstances that characterize a principal-agent interaction are drawn from a uniform distribution on $[0, 1]$. With enforcement ϕ, in all circumstances belonging to the interval $[0, \phi]$ opportunistic behavior would be detected, and the agent would be punished. We assume that the agent's payoff when punished is strictly lower than his payoff when acting honestly. This ensures that, whenever the specific circumstances surrounding the principal-agent interaction fall in $[0, \phi]$, behaving opportunistically is strictly dominated independently of the agent's type. A type O agent will therefore be willing to engage in opportunistic behavior only when the circumstances surrounding the interaction belong to $(\phi, 1]$. From an ex-ante perspective, this happens with probability $1 - \phi$.[9] In the buyer/seller example, a buyer matched with a type O seller is thus able to obtain the good with probability ϕ. The case where $\phi = 0$ corresponds to the scenario of null enforcement analyzed above. When $\phi \geq \alpha/\theta$, it is dominant to trust independently of the agent's type. Hence, the distribution of types within society is irrelevant for participation decisions. To make the problem relevant, we thus assume $\phi \in [0, \alpha/\theta)$ in the remainder

[8]The correlation between trustworthiness and trust is so strong that researchers find it hard to empirically identify the two variables. For instance, Glaeser et al. (2000) show that the answer to the World Value Survey question: "Generally speaking, would you say that most people can be trusted or you can't be too careful when dealing with people?" is a better predictor of an individual's *trustworthiness*, rather than of his/her *trusting behavior*.

[9]Our approach shares similarities with Tabellini (2008), where the quality of external enforcement is modelled by the probability of detection.

of the paper.[10]

Given ϕ, the expected net payoff for a type T principal is

$$E(U|x, \tau_P = T) = \left(\mu(x) + \frac{\sigma^2(x)}{\mu(x)}\right)\theta(1-\phi) + \theta\phi - \alpha \tag{12}$$

The equivalent for type O is

$$E(U|x, \tau_P = O) = \left(\mu(x) - \frac{\sigma^2(x)}{1-\mu(x)}\right)\theta(1-\phi) + \theta\phi - \alpha \tag{13}$$

From expressions (12) and (13), the expected net payoff from trusting is increasing in ϕ.[11] As one would expect, higher enforcement gives more incentive to participate to the principal. The difference between the net payoff from participation expected by a type T principal and that expected by a type O principal is

$$E(U|x, \tau_P = T) - E(U|x, \tau_P = O) = (1-\phi)\theta\frac{\sigma^2(x)}{\mu(x)(1-\mu(x))} > 0 \tag{14}$$

Notice that, relative to (9), an increase in ϕ reduces the difference in expected net payoffs. The next proposition summarizes the effects of external enforcement.

Proposition 2. *(Effects of external enforcement) Given a realization of the signal x, an increase in external enforcement (ϕ): (i) increases the expected net payoff from trusting for both types, (ii) reduces the difference between the net payoff expected by a trustworthy principal and the net payoff expected by an opportunistic principal.*

The intuition for point (i) is straightforward. The intuition for point (ii) is that, as enforcement increases, the agent's type becomes less important for the decision of whether to participate – since a principal may obtain the good with a positive probability even if she has the misfortune of being paired with an opportunistic agent. This has little effect on a trustworthy principal's incentives, since her attitude is a trusting attitude to start with. By contrast, the effect on opportunistic principals is much larger. With null enforcement, opportunistic principals are very reluctant to participate, since they tend

[10] We model enforcement with the exogenous parameter ϕ. Alternatively, one could view enforcement as a variable that is set by policy makers. So long as the policy makers' information consists in publicly available information, the analysis would be equivalent. If the policy maker possessed private information, then issues of signaling would arise. We leave this case to future research.

[11] The terms $\mu(x) + \frac{\sigma^2(x)}{\mu(x)}$ and $\mu(x) - \frac{\sigma^2(x)}{1-\mu(x)}$ represent the probabilities that a type T and a type O respectively attach to the agent being of type T. As a result, they are always less than one.

to project their own (opportunistic) type onto their agent. Overall, therefore, higher enforcement boosts the expected payoff from participation of an opportunistic principal more than it boosts that of a trustworthy principal. This weakens the selection effect.

To sum up, in this section we have accomplished two purposes. First, we have shown that different types of principals may rationally have different trusting attitudes – and may therefore be more or less inclined to participate as a result. Second, we have shown that introducing external enforcement boosts both types' expected net payoff from participation, although the effect is stronger for the opportunists. As a result of this latter effect, higher enforcement weakens the selection effect of introspection.

3 The evolution of trust and ethical attitudes

Point (i) in proposition 2 suggests that higher enforcement increases the expected return for both opportunistic and trustworthy types from participating in market interactions. Keeping everything else constant, this should then translate into higher market participation. There is, however, a caveat. Although in the short-run keeping the distribution of types within the population fixed may be appropriate, this is not the case if we consider a longer horizon. In the longer-run, the distribution of types within the populations is determined *endogenously,* as a result of a transmission process that may be affected by institutions. A full understanding of the relationship between institutions and individual behavior should take the long-term endogeneity of preferences into account. This is what we do in this section. We endogenize the share π of trustworthy and analyze the long run equilibria of the model. Our aim is that of characterizing the conditions under which the trustworthy can survive a process of natural (or, more to the point, cultural) selection in the long run, and whether this may depend on the level of enforcement.

At present, there is no universally accepted model of cultural evolution – indeed, in the words of Bowles (1998), "We know surprising little about how we come to have the preferences we do." For this reason, we abstract from a detailed analysis of the process of cultural transmission of traits, and instead adopt a reduced-form approach that borrows from evolutionary biology and evolutionary game theory. Differently from most evolutionary game theory, though, we do not analyze the evolution of *behavior*, but we consider rational behavior given the preferences and the beliefs associated (via introspection) with

the preferences. In our model, therefore, evolutionary forces operate on *preferences*. This is essentially the indirect evolutionary approach pioneered by Güth and Yaari (1992) and Güth (1995).[12]

3.1 Extended model

We start by adding some structure to the simple model presented in the previous section. We assume that there is a continuum $i \in [0, 1]$ of individuals. Individuals are either of type T or of type O and are identical in all other dimensions. Each individual is simultaneously involved in two interactions with two strangers (with whom she is randomly matched). In the first interaction, she acts as a principal and chooses whether to take part or not in an exchange. In the second interaction she takes the role of agent and chooses to behave honestly or opportunistically if trusted by the principal.[13] Information structure and payoffs are the same as in the previous section. Each individual i observes her own type and the (idiosyncratic) signal $x_i \in X$. Both the type and the realization of x_i are private information. Consistent with the evolutionary literature, we will interpret material payoffs as "fitness" in the rest of the paper.

3.2 Relative fitness and evolutionary stability

Let $X^\tau \subseteq X$ denote the set of signal realizations for which type τ chooses to participate. More precisely, X^T is the set of realizations of x such that (12) is positive, while X^O is the set such that (13) is positive. An individual i of type τ observing $x_i \in X^\tau$ will participate and obtain θ with probability $\pi + (1 - \pi)\phi$ and zero otherwise. The same individual observing $x_i \notin X^\tau$ will choose not to participate and will obtain α for sure. Consider now the payoffs that individuals obtain as agents. While type T individuals make zero profits,

[12] See also Bester and Güth (1998), Huck and Oechssler (1999), Samuelson (2004), and Samuelson and Swinkels (2006). Note that this approach (and in particular the replicator dynamic presented below) is consistent with a number of possible "micro-foundations", including Bisin's and Verdier's (2001) model of cultural transmission (see also Francois, 2008, for further details).

[13] This is meant to capture the notion that, in life, people are often simultaneously involved in a number of interactions, where they may play different roles. For instance, when someone buys a new house he is simultaneously a seller (for the old house) and a buyer (for the new house). Alternatively, we could have assumed that before exchange occurs, individuals are randomly selected to play the principal or the agent role, with probability 1/2 each.

type O obtain ρ with probability $1-\phi$, provided that they are matched with a principal who chooses to participate. Given the fraction of type T individuals in the population π, this happens with probability

$$\pi \int_{x \in X^T} dG(x|\pi) + (1-\pi) \int_{x \in X^O} dG(x|\pi), \tag{15}$$

where $\int_{x \in X^\tau} dG(x|\pi)$ is the fraction of type τ individuals who choose to participate. We can then write the difference between type T's and type O's average payoff as a function of the *actual* share of trustworthy individuals in the population,

$$\begin{aligned} \Omega(\pi;\phi) \quad = \quad & \left(\int_{x \in X^T} dG(x|\pi) - \int_{x \in X^O} dG(x|\pi) \right) (\pi\theta + (1-\pi)\phi\theta - \alpha) - \\ & \left(\pi \int_{x \in X^T} dG(x|\pi) + (1-\pi) \int_{x \in X^O} dG(x|\pi) \right) \rho(1-\phi). \end{aligned} \tag{16}$$

The first line of (16) is the difference in average material payoffs between trustworthy and opportunistic individuals in their role of principals. The second line represents the difference in their role of agents. Borrowing a term from evolutionary biology, we say that the quantity $\Omega(\pi;\phi)$ represents the *relative fitness* of type T given π. This is used below to characterize the evolutionary properties of different distributions of types in the population.

As shown by (16), the level of enforcement affects relative fitness both directly and indirectly. On the one hand, it lowers the probability that a principal is cheated by her agent. On the other hand, it may alter the individuals' propensity to participate. The latter effect emerges because different degrees of enforcement change the expected payoff from participation, as show in the previous section. This may affect X^T and X^O. If participation decisions are identical – namely, $X^T = X^O \neq \varnothing$ – then the first term in (16) is zero. In this case, relative fitness is always negative, owing to the opportunists' expropriation advantage. However, participation decisions need not be the same. From Proposition 1 we know that introspection makes trustworthy individuals more likely to participate. The selection effect thus implies that $X^O \subseteq X^T$. As a result, if π is sufficiently large so that

$$\theta(\pi + (1-\pi)\phi) > \alpha \tag{17}$$

then the first term of (16) is (weakly) positive. This is important, since it implies that if their share in the population is sufficiently high, the trustworthy may do better than the

opportunists as principals. As a result, when π sufficiently high, the sign of $\Omega(\pi;\phi)$ is not necessarily negative. [Indeed, in the next section we show that it can be positive.] By contrast, if the share of trustworthy in the population is so low that (17) does not hold, then opportunists do (weakly) better than trustworthy as principals, since they are less likely to participate. This is summarized in the next proposition.

Proposition 3. *(Relative fitness) From a material viewpoint, type O individuals always do strictly better than type T as agents. As principals, type O do (weakly) better (so that $\Omega(\pi;\phi) < 0$) for π sufficiently low. By converse, for π sufficiently high, type T individuals do (weakly) better, so that the sign of $\Omega(\pi;\phi)$ is ambiguous.*

Proposition 3 suggests that there are complementarities in trustworthiness, in the sense that being trustworthy rather than opportunist is more profitable when the trustworthy are majoritarian. In a way, trustworthy individuals "create their own space".[14] Intuitively, when their share in the population is low, the selection effect actually hurts trustworthy individuals, since trust is clearly misplaced. Things change when the share of trustworthy in the population is high. In this case, the advantage that trustworthy principals derive from the selection effect may overcome the opportunists' expropriation advantage as agents. There are two opposing forces at play. On the one hand, since the trustworthy are more likely to participate, the presence of many trustworthy individuals favors opportunistic agents, who are more likely to find gullible "victims".[15] On the other hand, the presence of many trustworthy individuals also increases the returns from participation – since the chances of meeting an opportunistic agent are remote. For relative fitness to be positive, the return from trusting an agent who behaves honestly (namely, θ) must be sufficiently high relative to the material payoff from opportunistic behavior (namely, ρ).

Evolutionary stability The equilibrium concept we use is that of *evolutionary stability*, first introduced by Maynard Smith and Price (1973). A trait $\tau = T, O$ is evolutionary stable if a population composed of individuals with the same trait τ cannot be successfully invaded by an alternative trait $\tau' \neq \tau$ that is initially rare. Hence, a state

[14]The expression is borrowed from De Long et al. (1990). They show that noise traders with an optimistic bias may obtain higher expected returns than rational arbitrageurs when enough individuals share the same bias. The mechanism behind of our effect is quite different, though.

[15]Given $X^O \subseteq X^T$, the second term in (16) is weakly increasing in π.

where all individuals are trustworthy (opportunists) is evolutionary stable if the average fitness obtained by trustworthy (opportunistic) individuals in this state exceeds that of the opportunists (trustworthy). Formally, for $\epsilon > 0$ vanishingly small, $\pi^* = 1$ and $\pi^* = 0$ are *evolutionary stable states* if $\Omega(1-\epsilon;\phi) > 0$ and $\Omega(\epsilon;\phi) < 0$, respectively. In some cases, the evolutionary process may not lead to homogeneous populations, but to a mixed population in which both type T and type O individuals coexist. We then say that $\pi^* \in (0,1)$ is a (*mixed*) evolutionary stable state if $\Omega(\pi^*;\phi) = 0$ and $|_{\pi=\pi^*} d\Omega/d\pi < 0$. These conditions are equivalent to requiring that π^* is asymptotically stable in the replicator dynamic (see Bowles, 2004, p. 72)

$$\pi' - \pi = \pi(1-\pi)\beta\Omega(\pi;\phi) \tag{18}$$

where π is the share of type T individuals in the current generation, π' is the share of type T in the next generation. The parameter $\beta > 0$ captures the speed with which the trait with higher fitness spreads among the population.[16]

Having introduced the notion of evolutionary stability within our framework, the next step is to characterize the evolutionary stable equilibria. Unfortunately, a full characterization of relative fitness in our framework is not possible without additional assumptions on the shape of the signal distribution, G. Inspection of (5) and (6) shows that expected payoffs are not necessarily monotonic in the realization of the signal x. While this is interesting from a theoretical viewpoint, it complicates the task of determining the sets X^T and X^O. For this reason, in what follows we restrict attention to the simple case of a binary signal, which guarantees monotonicity.

3.3 The binary signal case

Assume that $X = \{0,1\}$ and that $g(x=1|\pi) = \pi$ (so that $g(x=0|\pi) = 1-\pi$). In words, the probability of receiving the high signal ($x = 1$) when a fraction π of individuals are of type T is equal to π. Symmetrically, the probability of receiving the low signal ($x = 0$) is $1-\pi$. Denote also with $\Pi_n \equiv E(\pi^n)$ the n-th moment about the origin of $F(\pi)$. We

[16] The basic idea underlying the replicator dynamic is that individuals in the new generation tend to inherit the trait of their parents. However, a fraction of individuals in each generation will be exposed to "cultural models" different from their parents and may thus change their types. The probability of switching depends on the relative fitness, so that switching from type O to type T is more likely when $\Omega(\pi;\phi) > 0$, while the reverse happens when $\Omega(\pi;\phi) < 0$.

assume that the prior $F(\pi)$ does not change over time.[17] Given the information structure, the conditional expectation of π given $x \in \{0,1\}$ is

$$\mu(1) = \frac{\Pi_2}{\Pi_1},\ \mu(0) = \frac{\Pi_1 - \Pi_2}{1 - \Pi_1} \tag{19}$$

while the conditional variance is

$$\sigma^2(1) = \frac{\Pi_1\Pi_3 - \Pi_2^2}{\Pi_1^2},\ \sigma^2(0) = \frac{\Pi_2(1-\Pi_2) - \Pi_3(1-\Pi_1) + \Pi_1\Pi_2 - \Pi_1^2}{(1-\Pi_1)^2}. \tag{20}$$

Before stating the result, it is necessary to impose a weak requirement on the prior distribution $F(\pi)$. In Section 2, we assumed that the conditional variance $\sigma^2(x)$ was positive for all x. Given the conditional distribution of x, this has implication for the shape of the prior $F(\pi)$. The next assumption ensures that we restrict attention to priors that do not violate $\sigma^2(x) > 0$ for all $x \in \{0,1\}$.

Assumption 1. *The prior $F(\pi)$ is such that a) $\Pi_1\Pi_3 - \Pi_2^2 > 0$ and b) $\Pi_2(1-\Pi_2) - \Pi_3(1-\Pi_1) + \Pi_2\Pi_1 - \Pi_1^2 > 0$.*

Assumption 1 is satisfied for a broad class of priors, such as for instance the uniform distribution $(0,1)$ and, more generally, the whole class of Beta distributions. A full discussion of the technical implications of Assumption 1 is postponed to the next section. Given Assumption 1, it is immediate to verify that

$$R_2 \equiv \frac{\Pi_3}{\Pi_2} > R_1 \equiv \frac{\Pi_2 - \Pi_3}{\Pi_1 - \Pi_2} > R_0 \equiv \frac{\Pi_1 - 2\Pi_2 + \Pi_3}{1 - 2\Pi_1 + \Pi_2} \tag{21}$$

where R_0, R_1, and R_2 are obtained from (5) and (6). In particular, R_2 is the probability assessment that the agent is of type T made by a type T principal observing $x = 1$. R_1 is the same for a type T principal observing $x = 0$. The probability assessment of a type O who observes $x = 1$ is also equal to R_1. This is not surprising once we consider the fact that the conditional distribution of $x|\pi$ is identical to the distribution of $\tau|\pi$. Intuitively, the information of a type O who observes $x = 1$ is equivalent to the information of a type T who observes $x = 0$. Finally, R_0 is the probability assessment of a type O principal observing $x = 0$. Overall, therefore, (21) implies that an individual observing $x = 1$ expects the agent to be trustworthy with strictly higher probability than an individual observing $x = 0$. Moreover, given the same realization of x, a type T principal expects her agent to be trustworthy with strictly higher probability than a type O principal.

[17] Expectations are nonetheless affected by the dynamics of π through the signals x and τ.

From equations (12) and (13), a principal with assessment R_k, $k = 0, 1, 2$, will participate if[18]

$$R_k\theta(1-\phi) + \theta\phi - \alpha \geq 0 \tag{22}$$

or

$$R_k \geq \frac{\alpha - \theta\phi}{\theta(1-\phi)} \equiv R(\phi). \tag{23}$$

In words, for a given level of enforcement ϕ, a principal who believes that a randomly drawn agent is trustworthy with probability $R(\phi)$ would just be indifferent between participating or not. Note that $R(\phi)$ is strictly *decreasing* in ϕ and ranges between α/θ (when $\phi = 0$) and 0 (when $\phi = \alpha/\theta$).

The relationship between $R(\phi)$ and R_0, R_1 and R_2 determines the participation of the different types. In turn, this affects relative fitness, and determines which states may emerge in the long-run. For instance, if $R(\phi) > R_2$ then no individual (trustworthy or opportunistic) participates, independently of her signal's realization. Hence, $X^T = X^O = \varnothing$, and relative fitness $\Omega(\pi; \phi)$ is equal to zero for all π. In this case, there is no evolutionary stable state, but all $\pi \in [0, 1]$ are *neutrally stable*.[19] By contrast, if $R(\phi) \leq R_0$, then all individuals (trustworthy and opportunistic) participate, independently of the signal received. In this case, $X^T = X^O = X$. As a result, relative fitness $\Omega(\pi; \phi)$ is strictly negative for all π, so that $\pi = 0$ is the only evolutionary stable state. This situation generally arises when enforcement is sufficiently high to ensure that even the least trusting individuals in society (i.e., opportunists observing $x = 0$) choose to participate. Things become more complex when $R(\phi)$ takes intermediate values, i.e. it is located in the interval $(R_0, R_2]$. The next proposition provides a characterization of the evolutionary stable states in that case.

Proposition 4. *(Evolutionary stable states) Suppose that* $R(\phi) \in (R_0, R_2]$.

1. *If* $R(\phi) \geq (\theta - \rho)/\theta$, *then* $\pi = 0$ *is the only evolutionary stable state.*

2. *If* $R(\phi) < (\theta - \rho)/\theta$ *and*

[18] We adopt the convention that individuals participate when indifferent.

[19] Intuitively, while in the presence of a small shock π does not revert to its previous level, it does not move further away from it either.

(a) $R(\phi) \leq R_1$, *then* $\pi = 0$ *is evolutionary stable, but there may also exist (for parameters values) a mixed evolutionary stable state.*

(b) $R(\phi) > R_1$, *then* $\pi = 1$ *and* $\pi = 0$ *are both evolutionary stable (in fact, they are the only evolutionary stable states). The basin of attraction of* $\pi = 1$ *is* $(\theta R(\phi)/(\theta - \rho), 1)$.

Proof. See Appendix.

As proposition 4 highlights, a situation in which all individuals are of type T is evolutionary stable when the minimum probability that makes a principal willing to participate, $R(\phi)$, is less than $(\theta - \rho)/\theta$ and greater than R_1. The first requirement is necessary to generate complementarities in trustworthy ethical attitudes. For complementarities to emerge, the presence of a large share of trustworthy individuals in the population should benefit the trustworthy more than the opportunists. The benefit that the opportunists derive when there are many trustworthy about is that they are more likely to find gullible "victims" to expropriate. The benefit that the trustworthy derive from the presence of a large share of trustworthy is that expropriation is less likely. When $R(\phi) < (\theta - \rho)/\theta$, the first effect is weaker than the second, since the material benefits generated through the selection effect outweigh those generated by the expropriation advantage.[20]

The second requirement deals with participation. If $R(\phi) \leq R_1$, then even the opportunists tend to participate when the share of trustworthy is large. As a result, the selection effect has little bite.

Figure 1 illustrates the case in which both requirements are met.[21] When $R(\phi) > R_1$, the individuals who actually choose to participate are those who are trustworthy and have favorable information ($x = 1$). An opportunist would therefore not participate, even if she observed $x = 1$. If the material benefits generated by the selection effect are sufficiently large ($R(\phi) < (\theta - \rho)/\theta$), then, for π close to one, the opportunists achieve lower fitness than the trustworthy. Notice that in the equilibrium with $\pi = 1$ everyone participates (since the population is only composed of type T and all observe $x = 1$). The amount of

[20] The requirement $R(\phi) < (\theta - \rho)/\theta$ can equivalently be written as $\theta - \alpha > \rho(1 - \phi)$. The lhs captures the net gains from participation when $\pi = 1$ (the selection effect). The rhs captures the (expected) gains that can be reaped by opportunistic agents (the expropriation advantage).

[21] The Figure is based on a uniformly distributed prior. Parameters are as follows: $\theta = 1.5$, $\alpha = 1$, $\rho = 0.4$, $\phi = 0.2$.

investment is thus the maximum possible.

Figure 2 considers the case where $R(\phi)$ is greater than or equal to $(\theta - \rho)/\theta$. Here, the advantage afforded to type T through the selection effect is too small to overcome the opportunists' expropriation advantage. Hence, type T have lower fitness for all values of π. The unique evolutionary stable state is then $\pi = 0$. In this equilibrium, the population is totally composed of opportunists and all observe $x = 0$. Notice that, so long as $R(\phi) > R_0$ an opportunist observing $x = 0$ would not participate. As a result, the aggregate level of participation in the $\pi = 0$ equilibrium is zero.

Finally, Figure 3 deals with the case in which $R(\phi)$ is lower than $(\theta - \rho)/\theta$, but it is also lower than R_1. Since $R(\phi) \leq R_1$, an opportunist who has observed $x = 1$ would participate. Here, $\pi = 0$ is evolutionary stable, while $\pi = 1$ is unstable. However, a total takeover by type O is not the only possible outcome. For parameter values, there may exist a mixed evolutionary stable state. This is the case shown in Figure 3.[22] In the $\pi = 1$ equilibrium, overall participation is zero. In the mixed equilibrium, participation is partial, since opportunists observing $x = 0$ do not participate. Notice that, in this equilibrium, information effects prevent either type from spreading all the way. For instance, a marginal decrease in π would make $x = 1$ signals correspondingly less common. This would decrease participation by opportunists (but not by trustworthy), increasing relative fitness of type T. Similarly, a marginal increase in π would increase the opportunists' participation, which would lower relative fitness.

To sum up, when $R(\phi)$ takes intermediate values, a population composed only of opportunists is always evolutionary stable. However, other outcomes are possible. There are three alternative scenarios. In the first, a population of opportunistic individuals is the unique stable state. In the second, a population entirely composed of trustworthy individuals is also evolutionary stable. If this is the outcome, all individuals participate. In the third scenario, the trustworthy never completely take over. The population may settle in a mixed evolutionary stable state in which trustworthy and opportunists coexist.

Comparative Statics The conditions laid out in proposition 4 show that a necessary requirement for trustworthy attitudes to persist in the long-term is that the expropriation advantage enjoyed by opportunists should be lower than the material advantage

[22]The Figure is based on $\phi = 0.5$ and $\rho = 0.1$. Everything else is as in Figure 1.

that arises from greater market participation. Formally, this implies $R(\phi) < (\theta - \rho)/\theta$. Since $R(\phi)$ is increasing in α, a higher α makes the condition harder to satisfy. This is because a higher α increases the material payoff that can be reaped *without* participating to the market, which lowers the material advantage generated from the selection effect. Similarly, a higher ρ also makes the condition harder to satisfy, since it raises the expropriation advantage. Third, a higher value of θ increases the benefits reaped from the selection effect and therefore makes the condition easier to satisfy.

Finally, note that $R(\phi)$ is decreasing in ϕ. Changing the level of enforcement ϕ affects relative fitness in two ways. First, keeping participation behavior constant, a higher ϕ lowers the opportunists' expropriation advantage, and therefore raises relative fitness. Second, a higher ϕ lowers the risk of expropriation and therefore makes participation by opportunists more likely. This is the binary-signal equivalent of point (ii) in proposition 2.

External enforcement and ethical attitudes: *crowding in* and *crowding out*
We now look more closely at the relationship between external enforcement and ethical attitudes, by providing some formal results. Since the results are essentially implications of Proposition 4, they are presented as Corollaries.

Corollary 1a) (Crowding in) *Consider two enforcement levels ϕ' and $\phi'' > \phi'$. Then, $\pi = 1$ is evolutionary stable under ϕ'' but not under ϕ' if*

$$R_2 \geq R(\phi'') > R_1, \tag{24}$$

and either i) $R(\phi') < (\theta - \rho)/\theta$ and $R(\phi') > R_2$ or ii) $R(\phi') \geq (\theta - \rho)/\theta > R(\phi'')$.

Corollary 1b) (Crowding in) *Consider two enforcement levels ϕ' and $\phi'' > \phi'$. If both $R(\phi')$ and $R(\phi'')$ are such that $\pi = 1$ is evolutionary stable, then the equilibrium with $\pi = 1$ has a larger basin of attraction under ϕ'' than under ϕ'.*

Corollaries 1a) and 1b) illustrate how higher enforcement may actually *crowd in* trustworthy preferences in the long-run. This can happen in two ways. First, as in Corollary 1a), higher enforcement can make sure that opportunists cannot prosper in a society composed only of trustworthy. There are two types of situations where this may occur. One is where the selection effect would actually generate sufficiently high gains, but prior beliefs are too pessimistic for a selection effect to arise. Given the pessimistic prior, under low

enforcement the trustworthy would not participate, even after observing the high signal. In this case, a moderate increase in enforcement would induce trustworthy individuals with favorable information to participate, hence generating a selection effect. This is case (i) in the corollary. Case (ii) arises when the selection effect is too weak. This may for instance occur when the social gains generated by honest behavior by the agent $(\theta - \rho)$ are small, or when the payoff that a principal may obtain by not participating (α) is large. In this case, the selection effect may outweigh the opportunists' expropriation advantage only when enforcement exceeds some minimum threshold level. Hence, an increase in ϕ may induce convergence to the "good" equilibrium where all individuals are trustworthy.

Second, as shown in Corollary 1b), when both $\pi = 0$ and $\pi = 1$ are evolutionary stable, higher enforcement may expand the basin of attraction of the good equilibrium. An example is illustrated in Figure 4.[23] When enforcement is low $(\phi = \phi')$, the basin of attraction of the equilibrium with $\pi = 1$ is $(\pi', 1)$. By converse, when enforcement is high $(\phi = \phi'')$, the basin of attraction is $(\pi'', 1)$.[24] If the initial value of π lies between π'' and π', then, in the presence of low enforcement, π would over time converge to zero. In this case, a timely increase in ϕ (from ϕ' to ϕ'') may reverse the dynamics inducing convergence to the good equilibrium.

These results are broadly in line with the general idea of complementarity between institutions and social capital. However, greater external enforcement may also have unintended consequences in the long run. This is formalized below.

Corollary 2 (Crowding out) *Consider two enforcement levels ϕ' and $\phi'' > \phi'$. If $R(\phi') < (\theta - \rho)/\theta$, and*

$$R_2 \geq R(\phi') > R_1 \geq R(\phi'') > R_0, \tag{25}$$

then $\pi = 1$ is evolutionary stable under ϕ' but not under ϕ''. As a result, the long run level of participation may be lower under ϕ''than under ϕ'.

Higher enforcement may thus lead to an equilibrium with worse preferences – what we call *crowding out*. As a result of crowding out, overall participation may also fall. The effect behind crowding out is similar to that highlighted in Corollary 1a), but in reverse. Intuitively, with low enforcement $(\phi = \phi')$, only the trustworthy who observe

[23] In the Figure, $\phi' = 0.2$ and $\phi'' = 0.3$. Everything else is as in Figure 1.

[24] From Proposition 4, $\pi' = \theta R(\phi')/(\theta - \rho)$ and $\pi'' = \theta R(\phi'')/(\theta - \rho)$.

the high signal realization choose to participate.[25] If society is composed primarily of trustworthy people, then participating pays off, and the trustworthy may actually end up materially better off than opportunists. By contrast, if enforcement is high ($\phi = \phi''$), then opportunistic individuals observing the high signal also start to participate.[26] This eliminates the selection effect, since, when $\pi = 1$, an opportunistic "mutant" is as likely to participate as a trustworthy individual. The state $\pi = 1$ is therefore no longer evolutionary stable.

Since the equilibrium with $\pi = 1$ is characterized by the maximum level of participation, participation can only fall. If the fourth inequality in (25) holds – so that opportunists observing the low signal never invest – we can then have two scenarios.[27] In the first, the population converges to an equilibrium population comprising only opportunists. In this case, market participation collapses to zero. This is illustrated in Figure 5.[28] The second scenario arises when there is a mixed evolutionary stable state that prevents π from dropping to zero. In this state there is a positive share of opportunists observing the low signal who do not participate. Hence, despite a positive level of participation in equilibrium, participation is below the level achievable when $\pi = 1$ is evolutionary stable.

4 Interpretation and Robustness

Information imperfections play a major role in our analysis. We now discuss some issues of interpretations and robustness associated with our information structure.

In the previous section, the conditional distributions of the private signal x_i and of the individual's type were identical. In other words, the information gathered through introspection was as accurate as the information generated by the signal. A possible way to interpret this assumption is that each individual recalls a past event in which she has been able to observe the type of another individual randomly drawn from the population. For instance, she might have witnessed the behavior of someone playing as agent in another trust game.

[25] Formally, this is represented by the first and second inequality in (25).

[26] Formally, this is represented by the third inequality in (25).

[27] If it does not hold the equilibrium would still involve $\pi = 0$, but the level of participation would be unchanged, since even type O observing $x = 0$ would participate.

[28] In the Figure, we assumed $\phi' = 0.2$ and $\phi'' = 0.5$. Everything else is as in Figure 1.

In reality, adult individuals usually draw inferences from more than one past experience. The model can be generalized to accommodate individuals observing any finite number N of (conditionally on π) independent realizations of x. From a *qualitative* viewpoint, our results would not change. Clearly enough, a trustworthy individual observing N high realizations would expect the agent to be of type T with strictly higher probability than an opportunist observing the same vector of realizations. As a result, there exist a set of parameters values such that only a type T observing N high realizations would participate. This in turn implies that, for parameters values, $\pi = 1$ is evolutionary stable. On the other hand, it is clear that, as N increases, the beliefs of type T and type O individuals converge. In other words, introspection becomes less important. Hence, it is legitimate to ask whether, from a *quantitative* viewpoint, the effects emerging from our analysis are a reasonable approximation of real world effects. To a large extent, whether introspection matters for decisions is an empirical question, and the evidence says that it does. Butler, Giuliano, and Guiso (2009) show that not only do ethical attitudes explain trusting behavior, the relationship between trust and ethical attitudes persists even after principals have had the chance to collect information on the pool of potential agents.

There are also theoretical reasons to believe that introspection should matter more than a narrow reading of our analysis may suggest. First, we adopted a stripped down approach to model the trade off between the temptation to behave opportunistically and the psychological costs associated with cheating. We just assumed that the costs were sufficient to prevent opportunistic behavior in type T individuals. In reality, the temptation to cheat is likely to depend on what is at stake. This implies that the most informative past experiences for an individual are those in which the stakes were comparable to the problem in hand. For instance, I cannot infer from the fact that people are generally willing to help me when my car gets stuck, that I can entrust most people with my entire savings. On the other hand, individuals rarely have the possibility to experiment with high stakes. This suggests that we should expect introspection to become more relevant as the stakes become higher.

Second, as argued by Acemoglu, Chernozhukov and Yildiz (2008), the argument for the convergence of beliefs relies in part on simplifying assumptions of standard models of Bayesian learning. If trustworthy and opportunistic disagree on the way to interpret the information that they gather, asymptotic convergence is not assured. This implies that

introspection may still play a role even when individuals rely upon an arbitrarily large number of past experiences to take decisions.

There is also a deeper sense in which simplifying assumptions are relevant for interpretation. In our simple world, all relevant information about someone's ethical attitudes can be inferred from the outcome of the trust game he is playing. The real world is, of course, much more complicated. The outcomes of economic exchanges are affected by a large number of factors, besides parties' ethical attitudes. Observers typically only have partial information about these factors.[29] Hence, the fact that the information individuals possess is imprecise can be seen as a way to counterbalance our modelling of economic exchanges as extremely stylized interactions.

We now turn to Assumption 1. Since this is equivalent to assuming that the conditional variance $\sigma^2(x)$ is positive, Schwarz's inequality implies that both a) and b) in Assumption 1 are always satisfied with weak inequality.[30] Assumption 1 just imposes the strict inequality. This is satisfied for a broad class of priors. An example is the class of Beta distributions (including the uniform $(0,1)$). The only counter-example we could find in which the strict inequality does not hold is the case of a Bernoulli prior, which attaches positive probability only to $\pi = 1$ and $\pi = 0$. Since we find this case instructive, we discuss it briefly. The problem with the Bernoulli prior is that the posterior beliefs of a type T observing $x = 0$ (or a type O observing $x = 1$) are not well defined. This follows from the fact that observing $\tau = T$ and $x = 0$ (or $\tau = O$ and $x = 1$) is a zero probability event, given the prior. However, even in this extreme case, our main results would apply if we assumed that individuals use introspection when presented with zero probability events. In other words, all we need to assume is that, when the signal x is inconsistent with the individual's type, she does not believe that the signal is "more likely" to be correct than her type.[31]

To conclude, we address the broader issue of the evolutionary approach used in this paper. Observability of other players' preferences is generally considered necessary for

[29]Moreover, people typically lack a full "structural knowledge" of the structure of the game being played. Kurz (1994) shows that this may generate persistent heterogeneity of beliefs across individuals.

[30]Schwarz's inequality ensures that, if x and y are positive valued random variables, $E(xy)^2 \leq E(x^2)E(y^2)$. Setting $x = \pi^{1/2}$ and $y = \pi^{3/2}$ yields a). Setting $x = (1-\pi)^{1/2}$ and $y = \pi(1-\pi)^{1/2}$ yields b).

[31]In other words, the precision of the signal should not go to infinity faster than the precision of the type.

unselfish behavior to emerge spontaneously.[32] Our argument for the persistence of trustworthy behavior does not rely on the observability of other players' preferences. In this respect, we add a novel rationale to the existing literature. In our model, evolution solves the problem of inducing individuals to participate to the market – when this is optimal – via introspection. By giving people a preference for *trustworthiness*, it ensures that they generally trust others. However, evolution may operate on other dimensions as well. For instance, it could solve the same problem by giving people a direct preference for *trusting others*. For trustworthiness to persist in the long run, it is crucial that individuals are unable to develop "inconsistent" preferences. For instance, an opportunist with a direct preference for trusting others could destabilize the equilibrium where $\pi = 1$. Such an individual, however, would suffer from an irreconcilable conflict between her ethical attitudes/beliefs ("I only care about my material welfare and believe that most people, at the end of the day, do the same") and her trusting preferences ("I feel it is unfair to mistrust others"). The psychology literature on cognitive dissonance suggests that this type of conflicts are costly for the individual, causing anxiety, stress and other negative emotional states.[33] Individuals usually try to reduce conflicts by suppressing dissonant beliefs, attitudes, or behavior. This implies that the way our cognitive skills evolved may constrain our ability to develop inconsistent attitudes or to engage in inconsistent behavior.

5 Discussion and concluding remarks

Our analysis shows that the relationship between ethical attitudes and external incentives is quite complex. When external incentives (enforcement) are low, society may end up in a "bad" equilibrium, where opportunism is rife, and nobody complies. However, a "good"

[32] When preferences may be observed, Nash behavior may be temporarily destabilized by mutants who cooperate among themselves and defect with other agents. This is the idea behind the secret handshake model of Robson (1990). In contrast, when preferences are unobservable, evolutionary pressures should shape preferences so that individuals would behave "as if" they were playing Nash in a game in which payoffs represent the individual's fitness (see for instance Proposition 5 in Dekel, Ely, and Yilankaya, 2007, see also Samuelson, 2001, for a discussion of conceptual problems related to the observability of preferences).

[33] The theory was formulated by Festinger (1957). Aronson (1979) discusses experimental evidence. Applications to economics are developed by Akerlof and Dickens (1982).

equilibrium is also possible, where compliance rates are high. In this good equilibrium, all individuals are trustworthy, and compliance is motivated by internal norms of conduct, rather than by external incentives.

When external enforcement is high, good ethical attitudes may be "crowded out". At equilibrium, compliance is triggered exclusively by the threat of enforcement, and agents behave opportunistically whenever they can. The "good" equilibrium where ethical attitudes are trustworthy may no longer be possible. Overall market participation may also suffer as a result. Hence, our framework provides an example where high external enforcement may ultimately generate less market participation.

Our crowding out result shares similarities with Bohnet, Frey and Huck (2001) – henceforth BFH. These authors provide experimental evidence that supports the crowding out hypothesis. However, their theoretical explanation for the result is quite different from ours. In their model, introspection does not play any role. Rather, the result emerges because, as enforcement improves, principals are more willing to trust agents about whom they have unfavorable information. The fact that people possess individual-specific information about their counterparties' types plays a crucial role for the result. Here, we show that the crowding out effect also extends to the case of interactions among strangers. Moreover, in our model crowding out arises because higher enforcement encourages opportunistic principals to take more chances, and therefore counteracts the selection effect. The result is thus a consequence of the differences in the payoffs that different types are able to earn when acting *as principals*. By contrast, in BFH the result arises from the payoffs that different types obtain when acting *as agents*. The idea is that trustworthy agents are identified as such by principals, and therefore are trusted even when enforcement is low. This generates a direct material advantage for trustworthy agents. In our framework, this potential source of material advantage for trustworthy agents is ruled out by construction. The market works under conditions of complete anonymity, and the material payoff of agents behaving honestly is normalized to zero. Our explanation for crowding out can therefore be seen as complementary to that proposed by BFH.

A lesson that emerges from our analysis is that measures that are beneficial in the short-term may not necessarily be beneficial once their long-term effect on preferences is factored in. A policy that benefits opportunists relatively more than trustworthy would select in favor of opportunistic attitudes in the long-term. This would reinforce the very

behavior that the policy was set to counteract. In our framework, higher enforcement may end up doing just that, by encouraging greater participation by opportunists.

There is a general point here, which actually applies beyond the specifics of the model at hand. When assessing a policy, standard economic analysis usually concentrates on the policy's effects on the payoffs of different groups in society. Typical questions asked are: "Who benefits from the policy?" and "Who loses from it?". Crucially, if everyone in society would benefit from a policy, then the policy is generally deemed to be desirable. As we have seen, however, this approach may no longer be appropriate once the endogeneity of ethical attitudes is taken into account. In that case, comparing relative gains may become important. The question "Who benefits *relatively more* from the policy?" becomes crucial for understanding how the policy may affect the long-run distribution of ethical attitudes in society. This may for instance be relevant when considering "bail out" policies that forgive opportunistic behavior by key groups of individuals in society in the name of the common good.

Finally, although we have shown that higher external incentives may in some cases crowd out good ethical attitudes, crowding in is also possible. This may for instance occur when the gains that may be reaped from market participation are actually not very large. In this case, some minimum level of enforcement may be necessary in order to lure *anyone* to participate in the market. The presence of some external incentives may therefore provide the necessary leeway for trustworthy attitudes to spread within society.

6 Appendix

6.1 Proof of Proposition 4

We start with point 2 (b). Given $R(\phi) \in (R_1, R_2]$, only type T individuals who observe $x = 1$ invest. Hence, $X^T = \{1\}$ and $X^O = \varnothing$, which implies $\int_{x \in X^T} dG(x|\pi) = \pi$ and $\int_{x \in X^O} dG(x|\pi) = 0$. Rearranging the expression for relative fitness (16) one obtains,

$$\Omega(\pi; \phi) = \pi(\pi\theta + (1 - \pi)\phi\theta - \alpha - \pi\rho(1 - \phi)) \tag{6.1}$$

This can be rewritten as

$$\Omega(\pi; \phi) = \pi\theta(1 - \phi)\left[\pi\frac{\theta - \rho}{\theta} - R(\phi)\right] \tag{6.2}$$

If $(\theta-\rho)/\theta > R(\phi)$, then, for $\epsilon > 0$ vanishing, $\Omega(1-\epsilon;\phi) > 0$ and $\Omega(\epsilon;\phi) < 0$. Hence, both $\pi = 1$ and $\pi = 0$ are evolutionary stable. For $\pi = \hat{\pi} \equiv \theta R(\phi)/(\theta-\rho)$, $\Omega(\pi;\phi) = 0$. Since the derivative of Ω evaluated at $\pi = \hat{\pi}$ is positive, $\hat{\pi}$ is not evolutionary stable, but determines the basins of attraction of $\pi = 1$ and $\pi = 0$. Consider now point 2 (a). If $R(\phi) \in (R_0, R_1]$, type T invest for both signal realizations $\{0,1\}$. Type O invest when observing $x = 1$. Hence, $X^T = \{0,1\}$ and $X^O = \{1\}$, which imply $\int_{x\in X^T} dG(x|\pi) = 1$ and $\int_{x\in X^O} dG(x|\pi) = \pi$. From (16),

$$\Omega(\pi;\phi) = (1-\pi)(\pi\theta + (1-\pi)\phi\theta - \alpha) - [\pi + (1-\pi)\pi]\rho(1-\phi) \tag{6.3}$$

This can be rewritten as

$$\Omega(\pi;\phi) = (1-\pi)\theta(1-\phi)\left[\pi\frac{\theta-\rho}{\theta} - R(\phi)\right] - \pi\rho(1-\phi) \tag{6.4}$$

Notice that, for ϵ small, $\Omega(1-\epsilon;\phi) < 0$ and $\Omega(\epsilon;\phi) < 0$, which imply that $\pi = 0$ is evolutionary stable, while $\pi = 1$ is not evolutionary stable. Notice also that in this case (6.4) is an increasing-decreasing function of π which takes negative values at $\pi = 0$ and $\pi = 1$ and has an interior maximum at

$$\pi^M = \frac{1}{2} + \frac{\theta R(\phi) - \rho}{2(\theta-\rho)} \tag{6.5}$$

Substituting π^M in (6.4) shows that, if $\theta^2(1-R(\phi))^2 > 4\theta\rho - 4\rho^2$, then there exists a value $\pi^* \in (0,1)$ of π such that $\Omega(\pi^*;\phi) = 0$ and the derivative of Ω evaluated at π^* is negative. As a result, π^* is evolutionary stable.

Finally, consider point 1. Inspection of (6.2) and (6.4) shows that $\Omega(\pi;\phi)$ is negative for all $\pi \in [0,1]$ when $(\theta-\rho)/\theta \leq R(\phi)$. Hence, only $\pi = 0$ is evolutionary stable. □

References

[1] Acemoglu, D., Chernozhukov, V., and Yildiz, M. (2008) "Fragility of Asymptotic Agreement under Bayesian Learning" Mimeo.

[2] Adriani, F., and Sonderegger, S. (2009) "Why do Parents Socialize their Children to Behave Pro-Socially? An Information-Based Theory" CeFiMS Working Papers, DP96.

[3] Akerlof, G. A., and Dickens, W. T. (1982) "The Economic Consequences of Cognitive Dissonance" American Economic Review, 72: 307-319.

[4] Aronson, E. (1979) "The Social Animal", W.H. Freemam, San Francisco, CA.

[5] Bester H., and Güth, W. (1998) "Is altruism evolutionarily stable?" Journal of Economic Behavior and Organization 34: 193–209.

[6] Binmore, K. (1994) "Game Theory and the Social Contract. Volume 1: Playing Fair" MIT Press, Cambridge.

[7] Binmore, K. (2005) "Natural Justice" Oxford University Press, Oxford.

[8] Bisin, A., and Verdier, T., (2001) "The Economics of Cultural Transmission and the Dynamics of Preferences" Journal of Economic Theory, 97: 298-319.

[9] Bohnet, I., Frey, B.S., and Huck, S. (2001) "More Order with Less Law: On Contract Enforcement, Trust, and Crowding" American Political Science Review. 95: 131-144.

[10] Bowles, S. (1998) "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions" Journal of Economic Literature, 36: 75-111.

[11] Bowles, S. (2004) "Microeconomics: Behavior, Institutions, and Evolution" Princeton University Press, Princeton, NJ.

[12] Butler, J., Giuliano, P., and Guiso, L. (2009) "The Right Amount of Trust". Mimeo.

[13] Corneo, G, and Jeanne, O., (2009) "A Theory of Tolerance", Journal of Public Economics, forthcoming.

[14] Dawes, R.M. (1989)"Statistical criteria for establishing a truly false consensus effect" Journal of Experimental Social Psychology. 25: 1-17.

[15] Dekel, E., Ely, J.C., and Yilankaya, O. (2007) "Evolution of Preferences" Review of Economic Studies. 74: 685-704.

[16] De Long, J.B., Shleifer, A., Summers, L.H, and Waldmann, R.J., (1990) "Noise Trader Risk in Financial Markets" Journal of Political Economy, 98: 703-38.

[17] Ellingsen, T., and Johanneson M. (2004) " Promises, Threats and Fairness" Economic Journal, 114: 397–420.

[18] Festinger, L. (1957) "A Theory of Cognitive Dissonance", Row, Peterson, Evanston, IL.

[19] Francois, P. and Zabojnik, J. (2005) "Trust Social Capital and the Process of Economic Development" Journal of the European Economics Association, 31: 51-94.

[20] Francois, P. (2008) "Norms and Institutions", Mimeo.

[21] Glaeser, E.L., Laibson, D.I., Scheinkman, J.E., and Soutter, C.L. (2000) "Measuring Trust" Quarterly Journal of Economics, 115: 811-846.

[22] Guiso, L., Sapienza, P., and Zingales, L., (2008) "Trusting the Stock Market" Journal of Finance, 63: 2557-2600.

[23] Guiso, L., Sapienza, P., and Zingales, L. (2004) "The Role of Social Capital in Financial Development" American Economic Review, 94: 526-556.

[24] Güth, W. (1995) "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives" International Journal of Game Theory. 24: 323–44.

[25] Güth, W. and Yaari, M. (1992) "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game" in U. Witt. Explaining Process and Change – Approaches to Evolutionary Economics. Ann Arbor. 23–34.

[26] Hauk, E., and Saez-Martí, M., (2002), "On the Cultural Transmission of Corruption" Journal of Economic Theory 107: 311-335.

[27] Huck, S. (1998) "Trust, treason, and trials: An example of how the evolution of preferences can be driven by legal institutions" Journal of Law, Economics, and Organization 14: 44-60.

[28] Huck, S., and Oechssler, J. (1999) "The Indirect Evolutionary Approach to Explaining Fair Allocations." Games and Economic Behavior 28: 13-24.

[29] Kurz, M. (1994) "On the Structure and Diversity of Rational Beliefs," Economic Theory 4: 877-900.

[30] Maynard Smith, J., and Price, G., (1973) "The Logic of Animal Conflict", Nature 246: 15-18.

[31] Maynard Smith, J., (1982) "Evolution and the Theory of Games" Cambridge University Press, Cambridge.

[32] Orbell, J., and R.M. Dawes (1991) "A 'Cognitive Miser' Theory of Cooperators' Advantage" American Political Science Review. 85: 515-528.

[33] Rayo, L., and Becker, G.S. (2007) "Evolutionary Efficiency and Happiness" Journal of Political Economy, 115: 302-37.

[34] Robson, A.J. (1990) "Efficiency in Evolutionary Games: Darwin, Nash, and the Secret Handshake" Journal of Theoretical Biology 144: 379-396.

[35] Robson, A.J. (2001) "Why Would Nature Give Individuals Utility Functions?" Journal of Political Economy 109: 900-914.

[36] Ross L., Greene, D., and House, P., (1977) "The false consensus effect: An egocentric bias in social perception and attribution processes" Journal of Experimental Social Psychology 13: 279-301.

[37] Samuelson, L. (2001) "Introduction to the Evolution of Preferences" Journal of Economic Theory. 97: 225-230.

[38] Samuelson, L. and J. Swinkels (2006) "Information, Evolution and Utility" Theoretical Economics. 1:119-142.

[39] Samuelson, L. (2004) "Information-Based Relative Consumption Effects" Econometrica. 72: 93-118.

[40] Sapienza, P., Toldra, A., and Zingales, L., (2007). "Understanding Trust" CEPR Discussion Papers, 6462.

[41] Singer, T., and Fehr, E. (2005) "The Neuroeconomics of Mind Reading and Empathy", American Economic Review, Papers and Proceedings, 95: 340-345.

[42] Tabellini, G., (2008) "The Scope of Cooperation: Values and Incentives", Quarterly Journal of Economics, forthcoming.

[43] Vanberg, C., (2008) "A Short Note on the Rationality of the False Consensus Effect", Mimeo.

Figure 1- Both $\pi = 0$ and $\pi = 1$ are evolutionary stable

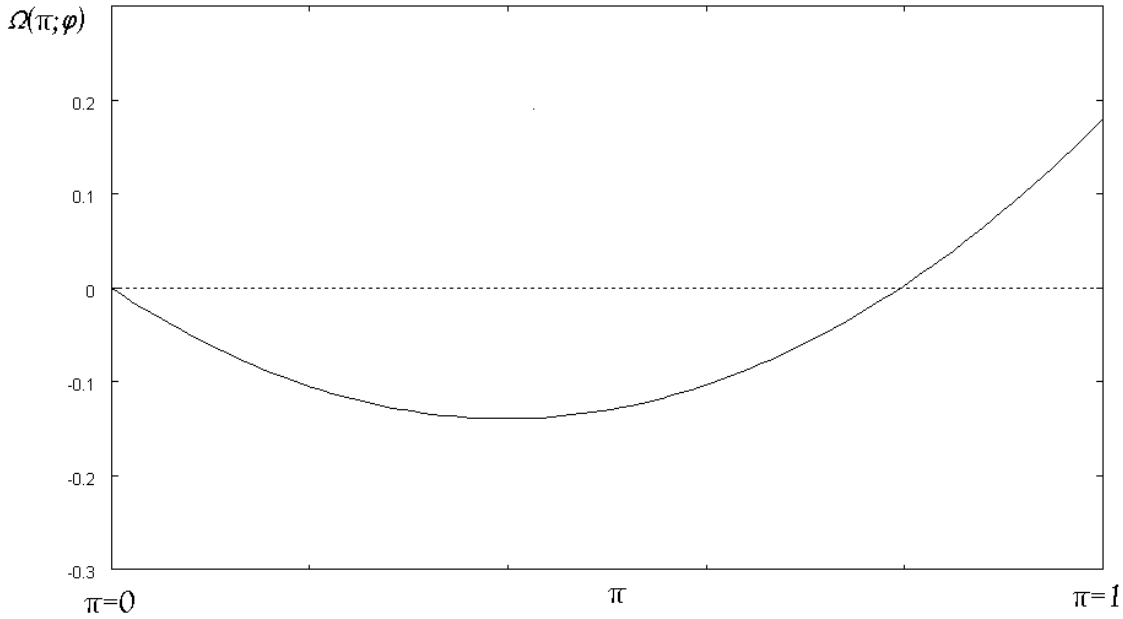


Figure 2 – Only $\pi = 0$ is evolutionary stable

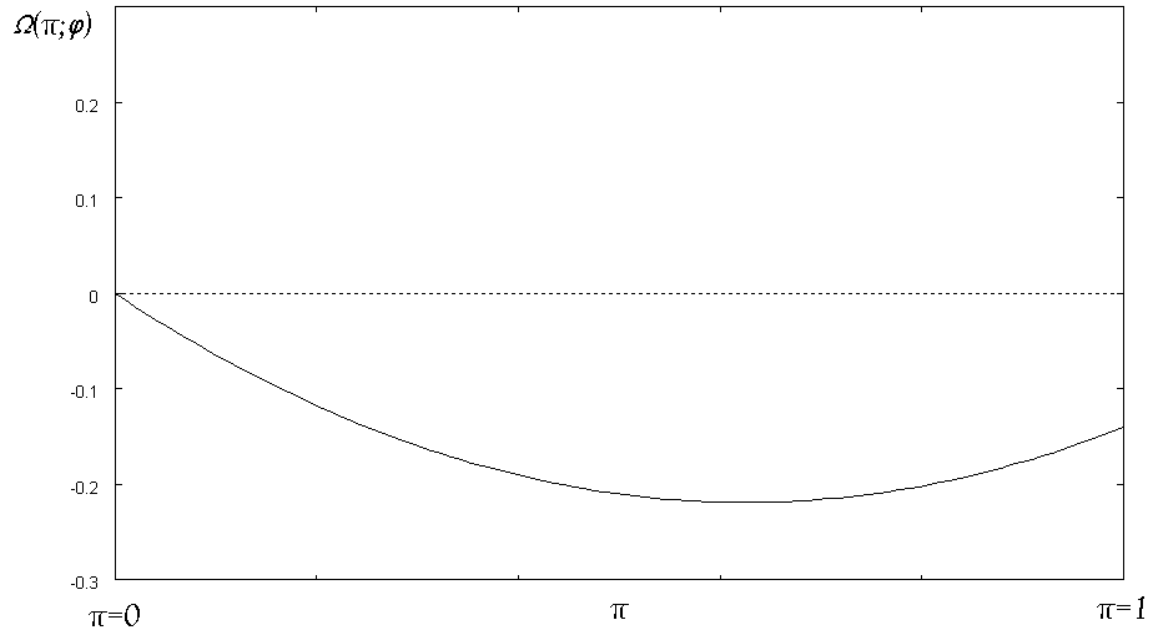


Figure 3 - $\pi = 0$ and $\pi^* < 1$ are evolutionary stable

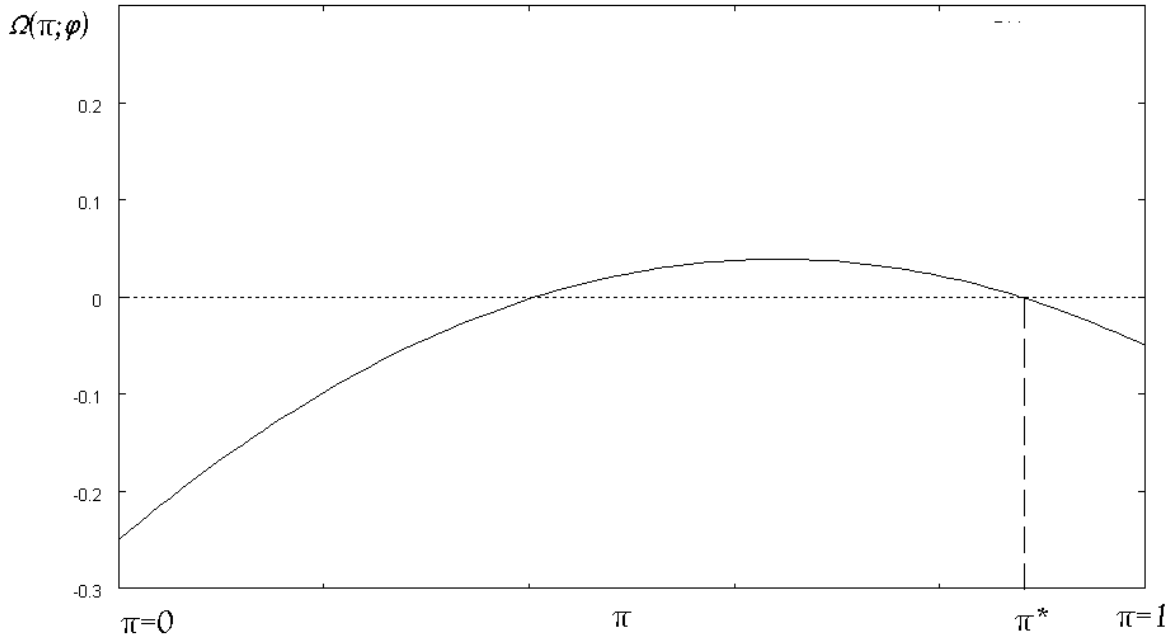


Figure 4 – A higher φ implies a larger basin of attraction of $\pi = 1$

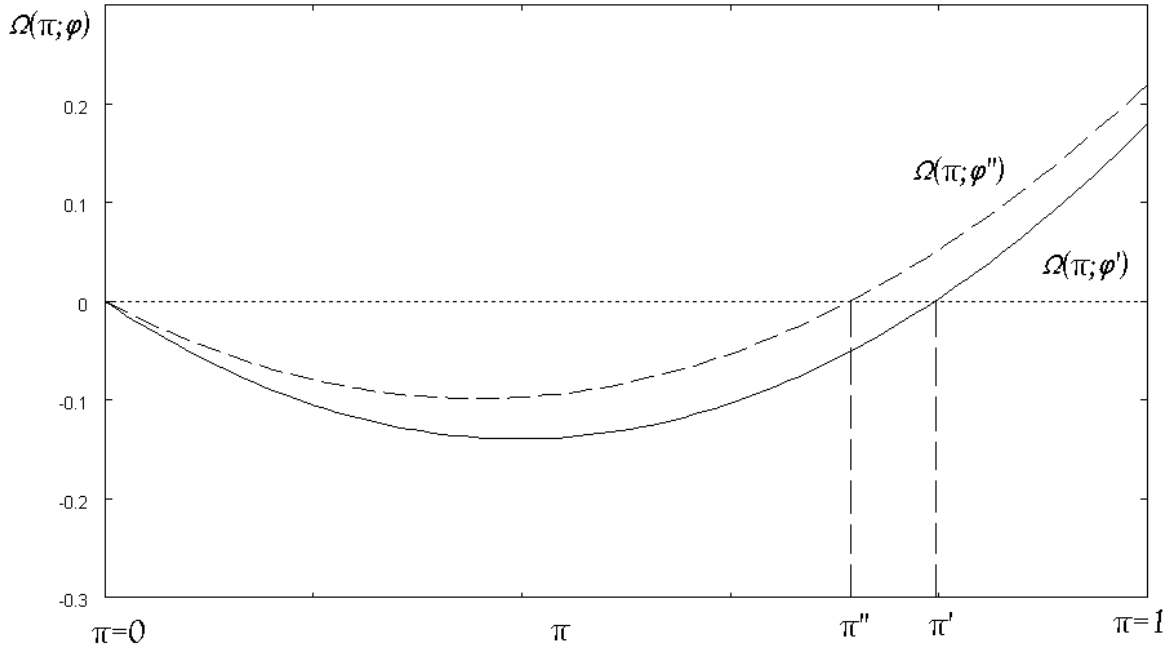
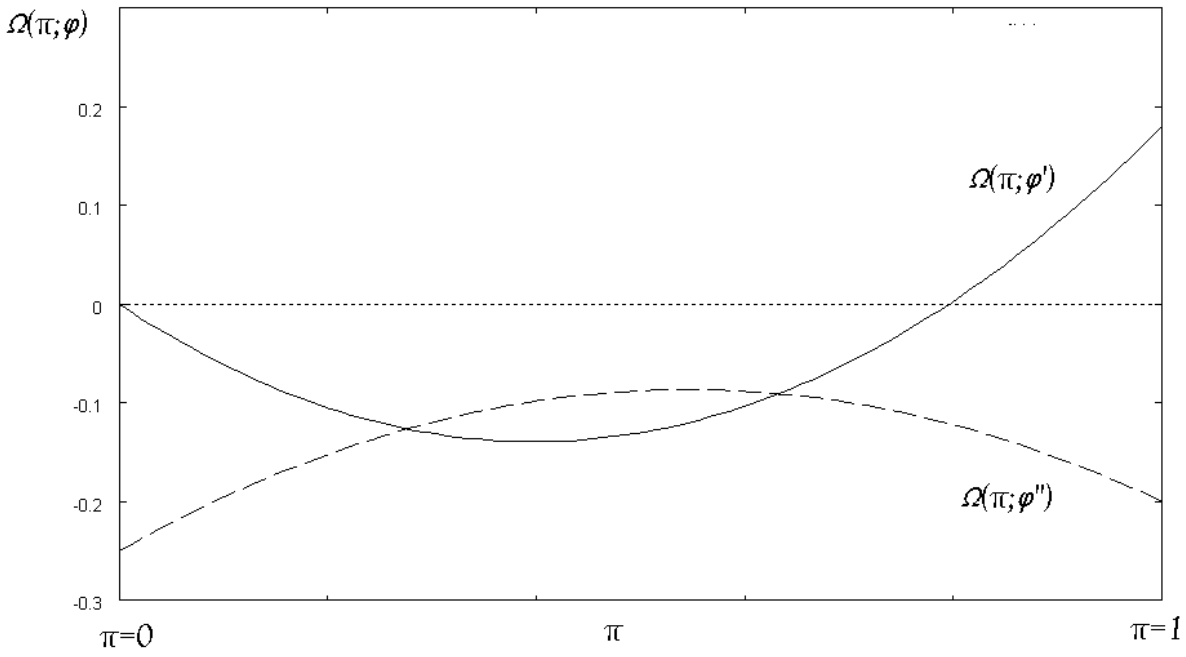


Figure 5 – Crowding out: a higher φ eliminates the stable equilibrium π = 1



7 Parameters and explanation for figures (material not meant for publication)

All figures are based on the assumption of uniform prior in $(0, 1)$. This implies

$$R_2 \equiv \frac{\Pi_3}{\Pi_2} = 0.75,\ R_1 \equiv \frac{\Pi_2 - \Pi_3}{\Pi_1 - \Pi_2} = 0.5,\ R_0 \equiv \frac{\Pi_1 - 2\Pi_2 + \Pi_3}{1 - 2\Pi_1 + \Pi_2} = 0.25. \tag{7.1}$$

All figures are based on $\theta = 1.5$ and $\alpha = 1$ and are generated by changing the values for ρ and ϕ.

Figure 1

In this case, $\rho = 0.4$ and $\phi = 0.2$. This implies $R(\phi) = 0.58\overline{3}$. Hence, $R(\phi)$ is between 0.5 and 0.75. As a result, only type T observing $x = 1$ invest. Hence, the share of type T who invest is π and the share of type O who invest is 0. From (16), relative fitness is

$$\Omega(\pi; \phi) = \pi\theta(1 - \phi)\left[\pi\frac{\theta - \rho}{\theta} - R(\phi)\right] \tag{7.2}$$

Figure 2

In this case, $\rho = 0.8$ and $\phi = 0.2$. $R(\phi)$ is still between 0.5 and 0.75 so that the expression for relative fitness is the same as in figure 1.

Figure 3

In this case, $\rho = 0.1$ and $\phi = 0.5$. Everything else is as in Figure 1. $\phi = 0.5$ now implies $R(\phi) = 0.\overline{3}$. Hence, $R(\phi)$ is between 0.25 and 0.5. As a result, all type T and the share of type O observing $x = 1$. The share of type T who invest is thus 1 and the share of type O is π. From (16), relative fitness is

$$\Omega(\pi; \phi) = (1 - \pi)\theta(1 - \phi)\left[\pi\frac{\theta - \rho}{\theta} - R(\phi)\right] - \pi\rho(1 - \phi) \tag{7.3}$$

Figure 4

In this case, $\rho = 0.4$ as in figure 1, $\phi' = 0.2$ and $\phi'' = 0.3$. $R(\phi') = 0.58\overline{3}$ while $R(\phi'') = 0.52381$. Both numbers are between 0.5 and 0.75. As a result, the expression for relative fitness is as in figure 1.

Figure 5

In this case, $\rho = 0.4$ as in figure 1, $\phi' = 0.2$ and $\phi'' = 0.5$. $R(\phi') = 0.58\overline{3}$, but $R(\phi'') = 0.\overline{3}$. Hence $R(\phi')$ is between 0.5 and 0.75, so that the expression for relative fitness is as in figure 1. $R(\phi'')$ is between 0.25 and 0.5 so that the expression for relative fitness is as in figure 3.