Adriani, Fabrizio and Sonderegger, Silvia (2015) Trust, trustworthiness and the consensus effect: an evolutionary approach. European Economic Review, 77 . pp. 102-116. ISSN 0014-2921

Trust, Trustworthiness and the Consensus Effect: An Evolutionary Approach*

Fabrizio Adriani†
University of Leicester

Silvia Sonderegger
University of Nottingham and CeDEx

April 1, 2015

*The paper has greatly benefited from comments by the editor, Jörg Oechssler, and two anonymous referees. We also thank Pierpaolo Battigalli, Simon Burgess, Ellen Greaves, Francesco Giovannoni, Luca Deidda, Bruno Frey, Simon Gächter, Luigi Guiso, Steffen Huck, Fabrizio Mattesini, Larry Samuelson, Joel Sobel, Robert Waldmann and seminar participants at various institutions and workshops for comments. We owe special thanks to Ken Binmore for discussions and encouragement.

†Corresponding author: Department of Economics, University of Leicester, University Road, Leicester, UK, LE1 7RH. fa148@le.ac.uk

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.

JEL CODES: A13, C73, D02, D03, D82, Z1.

KEYWORDS: Endogenous Preferences, Trust, Consensus Effect, Deterrence, Retribution, Crowding Out.

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 heterogenous) 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.

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 with most of the literature on the indirect evolutionary approach.[2] On the other hand, these models predict that those with unselfish preferences

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

[2]See e.g., Robson (1990). An exception to this is Huck and Oechssler (1998), who consider a setup

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, and 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,

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.

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

[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 and 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.

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.

The paper is organized as follows. In Section 2 we present the model. Section 3 characterizes the equilibrium in the stage game. In Section 4 we analyze the long term distribution of preferences. Section 5 focuses on the effects of deterrence and retribution. Section 6 briefly discusses possible extensions. Further details are provided in the working paper version.[7] Finally, Section 7 concludes by assessing the evidence on the relationship between trustworthiness, trust, and socioeconomic outcomes.

2 The Model

2.1 The stage game

Principals (P) We consider a sequential 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 (np), she will save her money and obtain a material welfare normalized to zero. If the principal chooses to participate (p) and the agent does not cheat (nc), the principal will obtain $\theta > 0$. In contrast, if the agent cheats (c), the principal suffers a loss $-\alpha$, with $\alpha > 0$. Hence, in this latter case, the principal would have been better off not participating in the exchange at all.[8] The *material* payoffs of the game are summarized in Figure 1. 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.

[7] Available at sites.google.com/site/fabrizioadriani/Home/research.

[8] Orbell and Dawes (1991) consider a sequential game where individuals choose whether or not to participate in a prisoners' dilemma game. The sequential game we use is simpler since the participation decision (trust/not) is built into the game. It is also clear that many of our results are not confined to the simple game we use. Other possible applications include the ultimatum game and the gift exchange game (see Section 7).

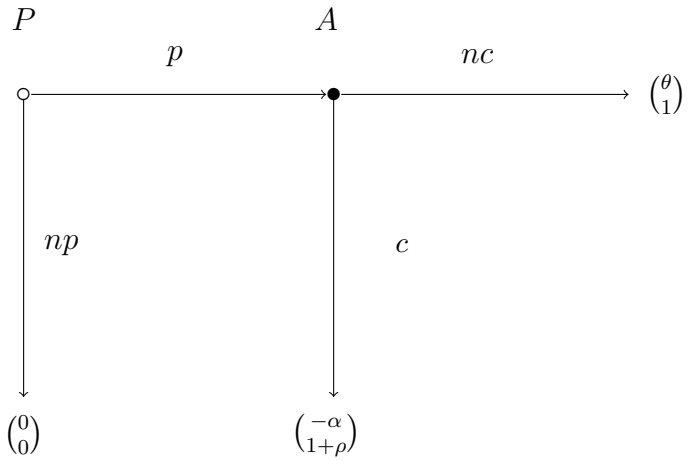

Figure 1: Material payoff game

Agents (A) If the principal chooses not to participate, the agent receives a material payoff equal to zero. When the principal participates, the agent obtains a payoff normalized to one if he does not cheat and a payoff equal to $1 + \rho$, $\rho > 0$, if he cheats. The agent's material welfare is thus maximized by cheating whenever trusted. We assume $\rho < \theta + \alpha$ – i.e. cheating is inefficient. In the buyer/seller example, the buyer may derive higher material welfare from consumption of the good, so that more surplus is generated if the good ends up in the buyer's hands rather than in those of the seller.

Preference Traits We assume that individuals may be of two types: Opportunistic (O) and Unselfish (U). Type O individuals only care about material welfare. In contrast with type O, type U individuals have other regarding preferences. We will focus here on the simplest possible form of other regarding preferences. In particular, we will assume that type U are altruistic and maximize the sum of their own material welfare and that of their counterparty. In the working paper version of this paper we show that our results are compatible with a variety of other regarding motives including reciprocity and homophily.[9]

[9]Empirical studies on deception (Gneezy, 2005) suggest that the propensity to cheat varies with the stakes. While we do not incorporate this effect in our model, it might provide an additional channel through which introspection may help unselfish individuals. Intuitively, an unselfish individual may be better equipped to figure out whether stakes are so high that even the unselfish agents may be tempted to cheat.

2.2 Information and beliefs

Following Dawes (1989), we consider an information structure where Bayesian learning leads to a consensus effect. We assume that it is common knowledge that all individuals (both principals and agents) are drawn from the same population, a continuum of size one. It is also common knowledge that the population contains both type U and type O individuals. However, the precise share of each type is unknown. This implies that, by looking at her own type, a principal can gather useful information about the likelihood that others (including the agent with whom she will be matched) are unselfish.

Formally, we denote with π the share of type U in the population (so that $1-\pi$ is the share of type O). Individuals have a common prior over π characterized by a non-degenerate cumulative distribution $F(\pi)$ and a density $f(\pi)$ with support $\mathcal{P} \subseteq [0,1]$. In addition to the prior, a principal has two pieces of relevant information: she observes a noisy signal $x \in X$ about π and her own type $\tau \in \{U, O\}$. The signal x captures the information that she is able to collect about the composition of the general population. Conditional on π, x has density $g(x|\pi)$ and cumulative $G(x|\pi)$, which are both continuous in π. We denote with $E(\pi|x)$ the expected value of π given the prior F and a realization x, and with $Var(\pi|x)$ the conditional variance (assumed positive for all $x \in X$). Notice that the signal x does not convey any agent-specific information. Our framework thus retains the assumption that preferences are unobservable. We will refer to the signal x as *objective evidence* in order to distinguish it from the type-dependent information that is generated by the observation of one's own type.

The timing of the stage game is as follows,

1. Nature matches each principal with an agent.
2. Players observe their own type τ. The principal also observes $x \in X$.
3. The principal chooses p or np.
4. The agent observes the principal's action and chooses c or nc.
5. Payoffs are realized.

2.3 The evolutionary process

Each generation is born with a share π of type U. Half of the population is randomly assigned to the role of principal and half to the role of agent. Individuals then play the stage game described above. After payoffs are obtained, a new generation (with possibly a different π) is born and its members are matched to play the game.

In order to pin down the long run distribution of preferences, we will focus on populations that are asymptotically stable under payoff monotone dynamics (see Samuelson and Zhang, 1992). This is a class of evolutionary dynamics comprising the standard replicator dynamics as a special case. Denote with $V_\tau(\pi)$ the average material payoff (across both roles) of type $\tau \in \{U, O\}$ given a share π of type U in the population. Under payoff monotone dynamics, π increases whenever $V_U(\pi) > V_O(\pi)$ and decreases when the reverse inequality holds. Notice that, since type τ's overall fitness is determined by the payoffs of both principals and agents, adaptation in our model works at the *population level* rather than at the *role level*.[10] The notion of asymptotic stability (see e.g. Weibull, 1997) reduces in our setting to the standard conditions

1. (Monomorphic populations) $\pi = 1$ ($\pi = 0$) is asymptotically stable if, for all $\epsilon > 0$ sufficiently small, $V_U(1-\epsilon) > V_O(1-\epsilon)$ ($V_O(\epsilon) > V_U(\epsilon)$)

2. (Polymorphic populations) an interior stationary point $\pi \in (0,1)$ (i.e. such that $V_U(\pi) = V_O(\pi)$) is asymptotically stable if $d(V_U(z) - V_O(z))/dz|_{z=\pi} < 0$

A monomorphism occurs if the population is entirely composed of individuals with one trait. Stability requires that rare mutants obtain lower fitness than the incumbent trait. A polymorphism arises when the two traits coexist ($\pi \in (0,1)$). In this case, both traits must have the same fitness and, after a small shock, the share of type U must revert to π.

In order to complete the description of the evolutionary process, we need to give an account of how beliefs change over time. As noted by relevant literature (see Robson and Samuelson, 2010, and Samuelson and Swinkels, 2006), it seems implausible that evolution could endow individuals with perfect priors. Once we accept that individuals are born

[10] Alternatively, we could have assumed that each individual is simultaneously involved in two interactions, playing in the role of principal in one and in the role of agent in the other. For instance, when someone buys a new house he is both a seller (for the old house) and a buyer (for the new house).

with imperfect information, we need to specify a mechanism through which beliefs about π adjust to changes in the actual value of π. There are two alternative ways to do this. The first would consist in assuming that individuals are born with a prior F which directly depends on the current value of π. This approach is however problematic since it implies that individuals need to know π (a parameter of the prior distribution) in order to estimate π. The second route, which we take, consists in specifying an indirect mechanism through which beliefs depend on the current value of π. Under this approach, individuals are born with the same prior F in every period but try to estimate π using the data available to them (i.e. the realization of the signal x and their own type). Notice that the fact that the in-built prior does not change with the actual value of π is merely semantic. One can always reinterpret the posterior distribution of π given x as the "relevant prior information" of the individual (i.e. all the information that is not generated by introspection). Since the distribution of x depends on the actual value of π, the distribution of (relevant) prior beliefs in the population depends stochastically on the actual share of unselfish individuals in the population.

2.4 Discussion

A central theme of our theory is that, through the consensus effect, the beliefs of individuals acting as principals depend on their social preferences. This induces correlation between the propensity to behave honestly and the propensity to participate. Hence, these propensities evolve together. In the standard (direct) evolutionary approach, evolution acts directly on behavior and a type is fully characterized by its strategy. Preferences and beliefs are redundant, so that there is no role for a consensus effect. Our analysis instead relies on the indirect evolutionary approach pioneered by Güth and Yaari (1992).[11] Under this approach, individuals (who can be Opportunists or Unselfish) select their behavior rationally given their preferences and the beliefs associated (via the consensus effect) with their preferences. In other words, we assume that individuals are not pre-programmed to adopt a certain behavior and rule out the possibility that evolution could somehow

[11]The evolution of preferences literature can be traced back to the work of Frank (1987). More recent contributions include, among others, Bester and Güth (1998), Huck and Oechssler (1999), Bisin and Verdier (2001), Samuelson (2004), and Samuelson and Swinkels (2006), Dekel et al. (2007), Alger and Weibull (forthcoming).

"hardwire" the optimal behavior, thus making the beliefs of the individual irrelevant.[12]

A key feature of our model sets it apart from most of the literature on the evolution of preferences. With a few exceptions (Huck and Oechssler, 1998, Gamba, 2013), most indirect evolutionary analyses assume that preferences are observable. Unselfish preferences are thus adaptive because they can affect the behavior of *other players*. This stands in contrast with our model, where one's preferences are private information. In our setting, unselfish preferences are adaptive because they shape the beliefs of the decision maker, thus affecting the decision maker's *own* behavior.[13]

3 Equilibrium play in the stage game

The mechanism whereby Bayesian learning generates a consensus effect hinges on the information conveyed by the observation of one's own type. In spite of the fact that individuals start with a common prior, posterior beliefs are type dependent. Formally, denote with $b(x, \tau_P) \equiv \Pr(\tau_A = U|x, \tau_P)$ the probability assessment that the agent is of type U made by a type τ_p principal who observes a signal realization x.

Lemma 1. *(The consensus effect) Given the same objective evidence, an unselfish principal assigns higher probability than an opportunistic principal to the agent being unselfish, i.e.*

$$b(x,U) = E(\pi|x) + \frac{Var(\pi|x)}{E(\pi|x)} > b(x,O) = E(\pi|x) - \frac{Var(\pi|x)}{1-E(\pi|x)},\ \forall x \in X. \qquad (1)$$

Proof. See Appendix.

The Lemma shows that, for any given value of x, the principal believes the agent to be unselfish with higher probability when she is herself unselfish. Individuals thus project their own characteristics onto others.

Consider now equilibrium play. A strategy for the principal maps her information $\{U, O\} \times X$ into a distribution over $\{p, np\}$. A strategy for the agent maps $\{U, O\} \times \{p, np\}$ into a distribution over $\{c, nc\}$. It is easy to verify that any sequential equilibrium of the game is such that type O agents play c whenever the principal participates. By contrast, since type U are altruistic and $\theta + \alpha > \rho$, type U agents play nc. Type O agents thus

[12]The working paper version provides a brief discussion of why it may not be a good idea to hardwire trusting behavior.

[13]We thank an anonymous referee for pointing this out.

enjoy an *expropriation advantage* since, by cheating, they maximize their material welfare, whereas type U "leave money on the table".

Consider now participation decisions.

Proposition 1. *Given the same objective evidence, the difference in the expected utility from participation between a type U and a type O principal is*

$$\underbrace{(\theta+\alpha)\frac{Var(\pi|x)}{E(\pi|x)(1-E(\pi|x))}}_{consensus\ effect} + \underbrace{\rho\left(1-E(\pi|x)-\frac{Var(\pi|x)}{E(\pi|x)}\right)+1}_{altruism} > 0. \tag{2}$$

Proof. See Appendix. □

Hence, type U principals have a higher propensity to participate than type O. As expression (2) suggests, there are two forces pushing in the same direction. First, the consensus effect implies that type U principals are more optimistic about the chances to be matched with a pro-social agent. This is captured by the first term in (2). The ratio $Var(\pi|x)/\left[E(\pi|x)(1-E(\pi|x))\right]$ has a natural interpretation. The numerator is a measure of the accuracy of objective evidence, x. The denominator is a measure of the accuracy of the other signal available to an individual, namely her type τ.[14] Overall, the higher the ratio, the more individuals will rely on introspection and thus the stronger the consensus effect. The second effect is a byproduct of altruism. Type U internalize their agent's material welfare when choosing whether to participate, while type O do not. This makes participation more attractive to type U. It is however worth emphasizing that the second effect is not crucial for our analysis. In the working paper version of this article we show that all our results carry through when we assume away all concerns for the agent's welfare by type U principals.

3.1 Coarse information

We now provide a detailed illustration of the consensus effect and its implications for participation for the simple case where the signal x is coarse. This case will be used extensively in the rest of the paper. Suppose that x can only take two values, i.e. $X = \{1, 0\}$ and that the probability of receiving the high signal is $g(x = 1|\pi) = \pi$ for all realizations of $\tau \in \{U, O\}$. Given the information structure, the signals x and τ are

[14] Using the law of total variance, it is easy to show that the denominator is equal to $Var(\boldsymbol{\tau} \mid x)$, where $\boldsymbol{\tau}$ is a random variable taking value 1 if $\tau = U$ and zero otherwise.

Conditional moments				
$E(\pi\|x=1)$	$=$	$E(\pi\|\tau=U)$	$=$	$\frac{\Pi_2}{\Pi_1}$
$E(\pi\|x=0)$	$=$	$E(\pi\|\tau=O)$	$=$	$\frac{\Pi_1-\Pi_2}{1-\Pi_1}$
$Var(\pi\|x=1)$	$=$	$Var(\pi\|\tau=U)$	$=$	$\frac{\Pi_1\Pi_3-\Pi_2^2}{\Pi_1^2}$
$Var(\pi\|x=0)$	$=$	$Var(\pi\|\tau=O)$	$=$	$\frac{\Pi_2(1-\Pi_2)-\Pi_3(1-\Pi_1)+\Pi_1\Pi_2-\Pi_1^2}{(1-\Pi_1)^2}$

Table 1: Conditional moments of π.

identically and independently distributed conditional on π. The conditional moments of π given either signal (x or τ) are reported for reference in Table 1, where $\Pi_n \equiv E(\pi^n)$ is the n-th moment about the origin of the prior $F(\pi)$.

Lemma 1 implies that

$$b(1,U) = \frac{\Pi_3}{\Pi_2} > b(1,O) = b(0,U) = \frac{\Pi_2 - \Pi_3}{\Pi_1 - \Pi_2} > b(0,O) = \frac{\Pi_1 - 2\Pi_2 + \Pi_3}{1 - 2\Pi_1 + \Pi_2}. \tag{3}$$

To see how this translates into behavior, suppose further that the prior $F(\pi)$ is uniform in $(0,1)$, so that $\Pi_1 = 1/2$, $\Pi_2 = 1/3$, and $\Pi_3 = 1/4$. Figure 2 depicts, for all values of θ, the locus of values for the probability that the agent is of type U that make a type τ principal indifferent between participating and not participating. A type τ principal observing the signal x will participate whenever $b(x,\tau)$ lies above the curve. Since type U principals internalize the welfare of their agents, their indifference locus lies below that of type O.

For θ sufficiently low, no participation occurs. As θ increases, type U observing the high signal choose to participate. Further increases in θ induce first type U principals with a low signal and then type O with a high signal to trust. Finally, for θ sufficiently large, all participate. Notice that, except for the extreme cases of full participation or no participation, the equilibrium participation rate depends on the actual share of unselfish individuals π. This happens for two reasons. First, type U are ceteris paribus more willing to participate. Second, the share of principals (of both types) observing the high signal realization depends on π.

4 The long run distribution of preferences

In this Section, we endogenize the share π of unselfish individuals and analyze populations that are asymptotically stable under payoff monotone dynamics.

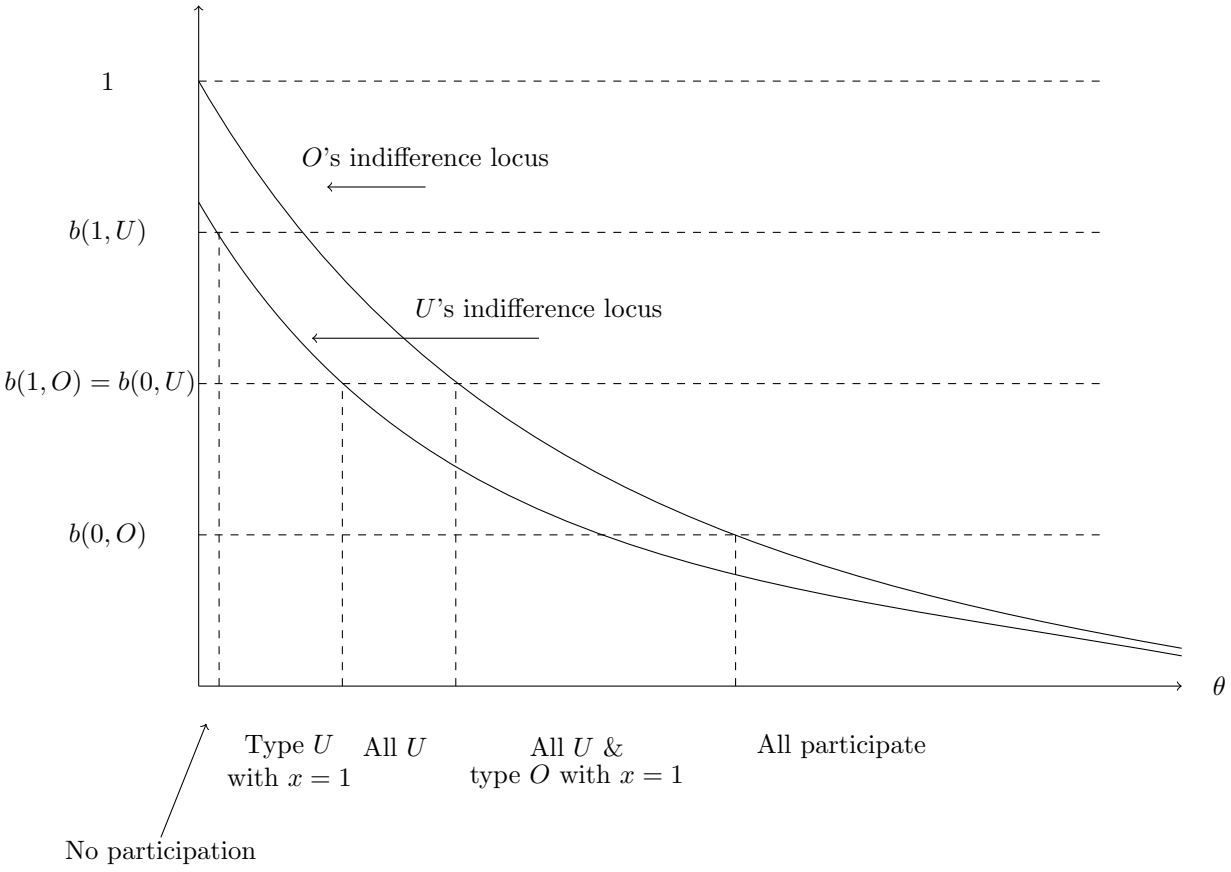

Figure 2: Participation decisions with uniform prior and binary signal.

Let $X^\tau \subseteq X$ denote the set of realizations of x for which type τ chooses to participate. A type τ principal observing $x \in X^\tau$ will participate and obtain θ with probability π and $-\alpha$ otherwise. The same individual observing $x \notin X^\tau$ will choose not to participate and will obtain zero for sure. As agents, type U individuals obtain a material payoff equal to one whenever trusted, so that the total average fitness of type U is

$$V_U(\pi) = \frac{1}{2}\mathcal{G}(X^U, \pi)[\pi(\theta + \alpha) - \alpha] + \frac{1}{2}\left[\pi\mathcal{G}(X^U, \pi) + (1 - \pi)\mathcal{G}(X^O, \pi)\right], \tag{4}$$

where $\mathcal{G}(X^\tau, \pi) \equiv \int_{x \in X^\tau} dG(x|\pi)$ represents the fraction of type $\tau \in \{U, O\}$ individuals who choose to participate given the actual share of unselfish individuals π. The last term in brackets is thus the probability of being matched, as an agent, with a principal who participates. The average fitness of type O is instead

$$V_O(\pi) = \frac{1}{2}\mathcal{G}(X^O, \pi)[\pi(\theta + \alpha) - \alpha] + \frac{1}{2}(1 + \rho)\left[\pi\mathcal{G}(X^U, \pi) + (1 - \pi)\mathcal{G}(X^O, \pi)\right]. \tag{5}$$

We can then write the relative fitness (i.e. the difference between type U's and type O's average fitness) as a function of the *actual* share of unselfish individuals in the population,

$$V_U(\pi) - V_O(\pi) = \frac{1}{2}\left(\mathcal{G}(X^U, \pi) - \mathcal{G}(X^O, \pi)\right)(\pi(\theta + \alpha - \rho) - \alpha) - \frac{1}{2}\mathcal{G}(X^O, \pi)\rho. \tag{6}$$

If participation decisions are identical – namely, $X^U = X^O$ – then the first term in (6) is zero. In this case, relative fitness is (weakly) negative for all π, owing to the Opportunists' expropriation advantage. The only candidate for asymptotic stability is thus a population entirely composed of Opportunists. However, participation decisions need not be the same. We know from Proposition 1 that unselfish individuals have a higher propensity to participate, i.e. $X^O \subseteq X^U$. If $\pi(\theta + \alpha - \rho) > \alpha$, then the first term of (6) is (weakly) positive. The sign of (6) is thus ambiguous and may change depending on π. This suggests that there may exist stable populations with a positive fraction of type U. We now provide a full characterization of the stable populations for the case of coarse information.

4.1 Stable populations with coarse information

The simple case of a binary signal $x \in X = \{0, 1\}$ introduced in Section 3.1 allows for a full characterization of the stable equilibria. We refer to the working paper version for the general case of a signal with $n \geq 2$ of realizations. All proofs for this subsection are

special cases of the slightly more general proofs given in Section 8.3 of the Appendix, and are therefore omitted.

If the population is entirely composed of Opportunists, participating is clearly suboptimal. Since greater propensity to participate is the only potential advantage of the Unselfish over the Opportunists, a population of Opportunists cannot be invaded. Formally, suppose that Π_2 and Π_3 in (3) are such that

$$b(1,U) \geq \frac{\alpha - 1 - \rho}{\theta + \alpha - \rho}. \tag{7}$$

Proposition 2. *When objective evidence is coarse, $\pi = 0$ is asymptotically stable iff (7) holds.*

Condition (7) ensures that type U would be willing to participate when observing the high signal. Since participation is suboptimal when the fraction of type U is sufficiently small, rare type U "mutants" have strictly lower average fitness than the incumbent type O. If condition (7) is violated, no one participates and both traits have identical fitness. In this case, $\pi = 0$ is neutrally stable but not asymptotically stable.

The next result provides necessary and sufficient conditions for an asymptotically stable polymorphic population where Unselfish and Opportunists coexist. Consider the following restrictions on material payoffs

$$\frac{\theta + \alpha}{\rho} > 2, \tag{8}$$

$$\theta - 2\sqrt{\rho(\theta + \alpha - \rho)} > 0, \tag{9}$$

and suppose that Π_1, Π_2, and Π_3 in (3) are such that

$$b(0,O) < \frac{\alpha}{\theta + \alpha} \leq b(1,O). \tag{10}$$

Proposition 3. *When objective evidence is coarse, conditions (8), (9), and (10) are necessary and sufficient for the existence of an asymptotically stable polymorphic population. Under these conditions the stable share of type U is*

$$\pi^* \equiv \frac{1}{2} + \frac{\alpha - \rho + \Delta}{2(\theta + \alpha - \rho)} \in (0,1), \tag{11}$$

where $\Delta \equiv \sqrt{\theta^2 - 4\rho(\theta + \alpha - \rho)}$.

Proposition 3 shows that, in the long run, different preferences may persist in the population. In order to gather intuition on how the polymorphic equilibrium may arise,

note that condition (10) ensures that type O principals participate only when they observe $x = 1$, while type U, who tend to be more optimistic, participate even when $x = 0$. Consider now what happens when π increases. This affects relative fitness in three ways. First, an increase in the share of unselfish principals means that the Opportunists are more likely to find gullible "victims" to expropriate. This reduces relative fitness. Second, an increase in π reduces the likelihood of being cheated. This benefits type U principals (who are more likely to participate) more than type O principals. This effect, which we call *participation effect*, generally increases relative fitness, thus inducing complementarity in unselfish preferences. The third effect is purely informational. An increase in π makes the high signal more common, thus increasing participation by the Opportunists. This weakens the participation effect. If π is large, so that participation is optimal, relative fitness is accordingly reduced.[15]

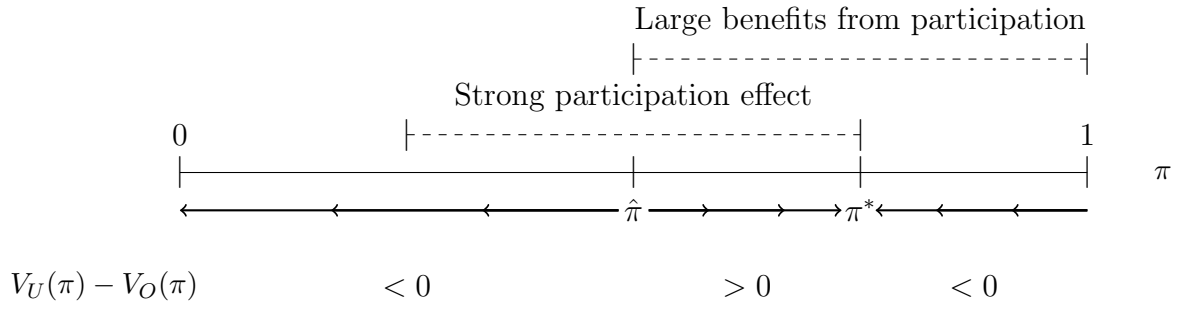

Figure 3: Stable polymorphic population.

The interplay of complementarities and substitutabilities generated by these three effects may generate a stable polymorphic population. This is illustrated in Figure 3. Below a critical value $\hat{\pi}$, relative fitness is negative. In this case, the expropriation advantage is the dominant effect. Moreover, when π is low, participation is suboptimal, so that unselfish principals are actually hurt by their higher propensity to participate. Relative fitness is also negative above the stable equilibrium π^*. Although participating is optimal in this case, most of the Opportunists observe the high signal. As a result, most Opportunists participate and the participation effect is weak. Moreover, further increases in π tend to favor the Opportunists more than the Unselfish. For intermediate values of $\pi \in (\hat{\pi}, \pi^*)$, the expected benefits from participation are sufficiently large and the partic-

[15]However, note that if π is low, an increase in the share of Opportunists who participate may actually increase relative fitness.

ipation effect is sufficiently strong to offset the expropriation advantage. Relative fitness is accordingly positive. It is then immediate to see why π^* is stable: Type U have higher fitness than type O for values of π immediately below π^*, while the reverse happens for values immediately above π^*. The existence of a range of values for π such that relative fitness is positive is ensured by conditions (8) and (9).[16]

Consider now the intuition for why $\alpha/(\theta+\alpha)$ has to belong to the interval $(b(0,O), b(1,O)]$ (condition 10). If $\alpha/(\theta+\alpha) \leq b(0,O)$, then even the Opportunists who observe the low signal $x = 0$ are willing to participate. As a result, there is no participation effect. If $\alpha/(\theta+\alpha) > b(1,O)$, then type O never participate. In this case, the expropriation advantage is the dominant effect for low values of π, while the participation effect dominates for sufficiently large values of π. As a result, any interior stationary point where the two effects offset each other is unstable.[17]

We now provide a simple numerical example of a polymorphic equilibrium.

Example Consider the example presented in Section 3.1 with uniform prior. Given a uniform prior, $b(0,O) = 1/4$ and $b(1,0) = 1/2$. Hence, for (10) to be satisfied, the ratio θ/α must be between 1 and 3. If this holds, type O participate only when observing the high signal, while type U always participate. Assume then the following parameter values: $\theta = 5$, $\alpha = 2$, and $\rho = 1$. It is immediate to check that conditions (8-9) hold, so that there is a stable polymorphic population where the share of type U is $\pi^* = 2/3$, i.e. two thirds of the population are unselfish, with a third of Opportunists. Given $\pi^* = 2/3$, the average payoff in the population is $1.\overline{7}$.[18]

[16]The requirement $(\theta+\alpha)/\rho > 2$ in (8) can be equivalently rewritten as $\theta+\alpha-\rho > \rho$. The term $\theta+\alpha-\rho$, which appears both in (8) and in (9), represents the joint net surplus created by not cheating. Inspection of (8) thus suggests that type U only survive when the net externality given to principals by type U agents is large enough. However, (9) points to a countervailing effect. Since type U principals benefit relatively more from the externality and given that fitness must be identical for both types, a larger $\theta+\alpha-\rho$ must be offset in equilibrium by a higher probability to be cheated (i.e. a lower π^*). However, through the information channel, a drop in π^* inhibits participation by type O. If the net externality becomes too large, a drop in π^* further widens the gap between type U's and type O's fitness, thus making any interior stationary point unstable.

[17]In this case, a monomorphic population of type U may be stable. See below.

[18]The average payoff is given by

$$\frac{1}{2}[\pi^* + \pi^*(1-\pi^*)][\theta\pi^* - \alpha(1-\pi^*) + \rho(1-\pi^*) + 1],$$

where $\pi^* + \pi^*(1-\pi^*) = 0.\overline{8}$ is the share of the population who participate, $\theta\pi^* = 3.\overline{3}$ is the average payoff

Finally, for given configurations of the parameters, there also exist stable monomorphic populations entirely composed of unselfish. It is straightforward to verify that this happens whenever $b(1,O) < \alpha/(\theta+\alpha) \leq b(1,U)$ and $\theta > \rho$. These equilibria do not appear particularly plausible, though. They occur when the consensus effect is so strong that a rare type O mutant in a population entirely composed of unselfish individuals would use introspection and choose not to trust even when objective evidence suggests otherwise.[19]

5 Implications for punishment and prevention

We consider here an extension where, in a fraction $\phi \in (0,1)$ of interactions, cheaters are detected and punished, i.e. rather than obtaining $1+\rho > 0$, a cheater receives a material payoff $1-\lambda$, with $\lambda > 0$. We will use two polar cases to fix ideas on the roles performed by punishment. In the first, potential cheaters are perfectly able to anticipate punishment and will not cheat unless they are sure that they can get away with it. In this case, no actual punishment is enforced in equilibrium, so that punishment only takes the form of *deterrence*. In the second case, potential cheaters cannot anticipate whether they will face punishment for their misbehavior and retain the incentive to cheat in all interactions (provided λ is not too large). As a result, punishment is ineffective as deterrent and only takes the form of ex-post *retribution*.

Deterrence Suppose then that agents know whether they will be punished or not in case of misbehavior. If principals could also perfectly anticipate punishment, then it is easy to show that deterrence would have no consequence on the long run distribution of preferences.[20] We therefore focus on the more realistic case where, at the time the trusting decision is taken, the principal is imperfectly informed. In particular, he does not know whether a cheating agent would be punished or not, he only knows that opportunistic

from participating, $\rho(1-\pi^*) = 0.\overline{3}$ is the average payoff from cheating times the share of the population who cheat, and $\alpha(1-\pi^*) = 0.\overline{6}$ is the loss from being cheated times the proportion of participating principals who are cheated.

[19] In the working paper version of this manuscript, we discuss these equilibria more extensively. We also argue that they disappear if we endogenize the weight that individuals assign to introspection vis-á-vis objective evidence when forming their posterior beliefs.

[20] Intuitively, in the fraction ϕ of interactions where punishment could be enforced, social preferences play no role. In the rest of the interactions, the problem would be identical to the case of no punishment.

agents cheat with probability $1-\phi$. The case where $\phi = 0$ corresponds to the scenario of no enforcement analyzed above. Denote with $R(\tau;\phi)$ the critical value of the probability that the agent is of type U which makes a type τ principal indifferent between participating and staying out. This is given by[21]

$$R(\tau;\phi) \equiv \begin{cases} \frac{(1-\phi)\alpha-\phi\theta}{(1-\phi)(\theta+\alpha)} & \tau = O \\ \frac{(1-\phi)(\alpha-\rho)-1-\phi\theta}{(1-\phi)(\theta+\alpha-\rho)} & \tau = U \end{cases}. \tag{12}$$

Similar to the case of no punishment, $R(U;\phi) < R(O;\phi)$ for all ϕ. i.e. type U are ceteris paribus more willing to participate. Notice also that $R(\tau;\phi)$ is strictly *decreasing* in ϕ. As one would expect, better enforcement increases the propensity to participate of both types. In the short term, deterrence thus raises aggregate material welfare both by reducing the scope for cheating and by encouraging participation.

On the other hand, deterrence makes participation decisions less type-dependent. Simple algebra shows that the first term in (2) becomes

$$(1-\phi)(\theta+\alpha)\frac{Var(\pi|x)}{E(\pi|x)(1-E(\pi|x))}. \tag{13}$$

Compared to (2), a positive ϕ thus reduces the importance of the consensus effect for participation decisions. Intuitively, as enforcement becomes more effective, one's expectations about her counterparty's type become less important for the decision of whether to participate – since dishonest behavior may be prevented even if one has the misfortune of being paired with an opportunistic agent.

In order to understand what this implies for the long term distribution of preferences, we provide a graphical illustration of the complex interaction between deterrence and the share of type U. Details are provided in the Appendix.[22]

Figure 4 shows the relationships between ϕ and the stable share of type U (top part), and between ϕ and material welfare (bottom part). Starting from $\phi = 0$ and with a

[21] The expected utility from participation for a type τ individual observing signal x is

$$\begin{aligned} E(w|x,O) &= (\theta+\alpha)[b(x,O)(1-\phi)+\phi]-\alpha, \\ E(w|x,U) &= (\theta+\alpha-\rho)[b(x,U)(1-\phi)+\phi]+\rho+1-\alpha. \end{aligned}$$

Setting the above equal to zero and solving for the b terms yields $R(\tau;\phi)$.

[22] In particular, Section 8.3 provides a generalization of Propositions 2 and 3 for the case $\phi \geq 0$. There, we also make precise the conditions under which deterrence may crowd in or crowd out other regarding preferences.

set of parameters which induce $\pi = 0$ as the unique stable equilibrium, an increase in ϕ makes participation more profitable and reduces type O's expropriation advantage. This has two effects. First, by increasing type U's fitness vis-à-vis type O's, a higher ϕ makes a polymorphic equilibrium viable.[23] Second, a higher ϕ implies a higher equilibrium share of type U. The intuition for this result is that, since the two types must have the same fitness in equilibrium, the equilibrium proportion of type O who participate must increase to compensate for the lower expropriation advantage. For this to happen, a higher proportion of type O must observe the high signal, so that the share of type U needs to increase accordingly.

However, closer inspection of Figure 4 reveals that there is also another, more subtle effect at work. As ϕ increases further – so that it exceeds another threshold given by the indifference condition for type O principals with $x = 0$, $R(O; \phi) = b(0, O)$ – the unique stable population is again $\pi = 0$. More effective deterrence may thus have a *crowding out* effect. Intuitively, when there is little deterrence, only the Opportunists who observe the high signal realization choose to participate (while *all* the Unselfish participate). A polymorphic equilibrium is thus possible. By contrast, when deterrence is more widespread, the Opportunists observing the low signal also choose to participate. Since the Opportunists are now as likely to participate as the Unselfish, no stable polymorphic population exists. In the working paper, we present parameter values such that better enforcement, by eliminating the polymorphic equilibrium, ultimately leads to lower welfare.

Retribution Consider now the other extreme, in which agents are unable to anticipate punishment and deterrence is absent. How does retribution affect payoffs? In our model, the effect of retribution is essentially that of reducing the expected payoff achieved by cheating. Define $\tilde{\rho} \equiv (1 - \phi)\rho - \phi\lambda$ as the expected net return from cheating when the agent only knows that cheating will be punished with probability ϕ. In order to focus on pure retribution, we assume that $\phi \in [0, \rho/(\rho + \lambda)]$, so that $\tilde{\rho} \geq 0$. If $\tilde{\rho}$ were negative, then cheating would become unprofitable and thus no cheating would occur, i.e. we would have achieved full deterrence. It is clear that, once ρ is replaced with $\tilde{\rho}$, the model with

[23] For parameter values, a stable polymorphic equilibrium is only possible if ϕ is above a certain threshold. This is implicitly given by the value of ϕ, ϕ^*, satisfying,

$$R(O; \phi^*) = \frac{\theta + \alpha - 2\sqrt{\rho(\theta + \alpha - \rho)}}{\theta + \alpha},$$

As shown in the Appendix, $\phi \geq \max\{\phi^*, 0\}$ is necessary for a stable polymorphic equilibrium.

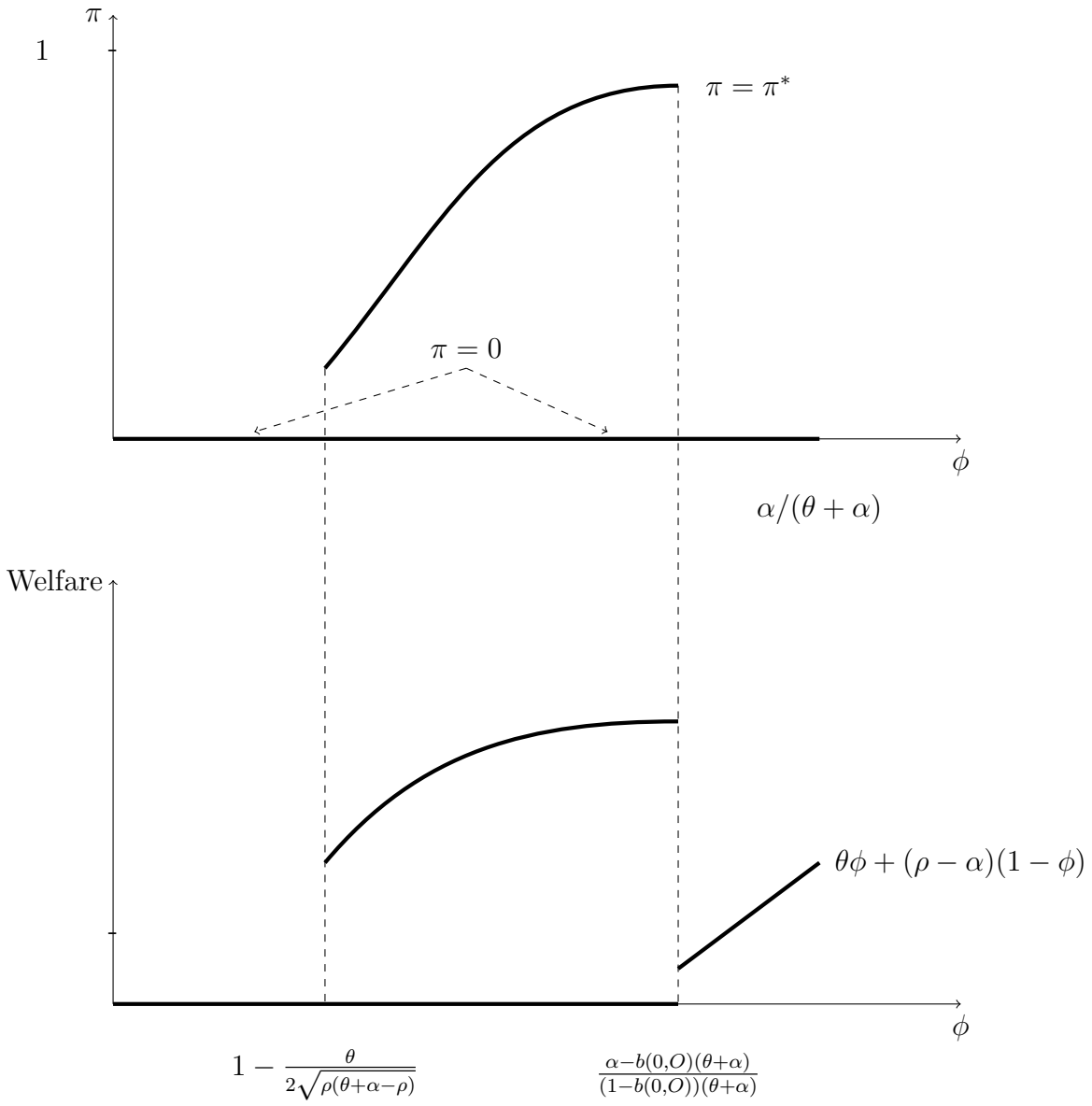

Figure 4: Share of type U and average welfare as a function of ϕ.

retribution is isomorphic to the one seen in the previous sections. As long as $\tilde{\rho}$ stays positive, an increase in ϕ always reduces welfare in the short run. That is, the cheaters obtain on average a lower payoff, whilst the payoff of the others remain unaffected. On the other hand, in the long run, a lower $\tilde{\rho}$ makes the necessary conditions for a polymorphic equilibrium (8) and (9) easier to satisfy. Moreover, the share of unselfish individuals in the polymorphic equilibrium is strictly increasing in $\tilde{\rho}$. This suggests that retribution may have positive effects on the distribution of preferences in the long run. Different from the case of deterrence, however, crowding out is not possible. This is because, by construction, retribution does not affect participation decisions.

To sum up, while deterrence increases welfare in the short term (when the distribution of types is fixed), it has ambiguous effects in the long run. This is because, as we have shown, deterrence weakens the participation effect. In turn, this may crowd out other regarding preferences, leading over time to more cheating and lower welfare.[24] Retribution has a somewhat opposite effects. While it is always wasteful in the short run, it induces a more desirable distribution of preferences in the long run.

6 Robustness and extensions

In the working paper version of this paper, we extend the model in the following directions

1. we consider the case where the principal, prior to her decision, can sample a number n of individuals and observe their preferences, so that the signal x takes a finite number n of realizations;

2. we consider other regarding preferences different from altruism;

3. we allow for individuals who are not constrained by Bayes rule over the weight to give to introspection via-à-vis objective evidence when forming their beliefs.

In relation to 1, we find that, given any finite n, there exist parameters that support a polymorphic equilibrium. However, the set of suitable parameters tend to shrink in size as n becomes larger. Regarding point 2, we show that altruism is not crucial for our findings. In particular, we obtain identical results with reciprocity. We also show that

[24] Bohnet, Frey and Huck (2001) provide evidence of better enforcement crowding out trustworthiness. Their concept of enforcement has both elements of deterrence and retribution.

the same results apply even when type U individuals have selfish preferences when acting as principals, so that the participation effect is purely driven by introspection. As for point 3, we argue that allowing evolution to shape the way in which individuals combine different sources of information would eliminate any monomorphic equilibrium where the population is entirely composed of type U. Polymorphic equilibria are however robust to the introduction of these mutants.

7 Empirical evidence and concluding remarks

Our work builds on the consensus effect to derive implications for the relative material gains of opportunistic and unselfish individuals. The importance of the consensus effect for trust is amply documented in the experimental literature.[25] We focus here on the implications of our theory for the long run distribution of social preferences. In our model, one's propensity to trust is determined by his trustworthiness. Trustworthiness and opportunism are then passed on to future generations under payoff monotonic dynamics. A possible interpretation of our model is thus that trustworthiness is genetically determined and the most successful trait produces a larger number of surviving offspring. An alternative interpretation is that youngsters have a propensity to adopt the cultural traits of successful individuals of the previous generation. Under both interpretations, a key prediction of our model is thus that trustworthy individuals can be as successful as the opportunists. Butler et al. (forthcoming) present evidence that supports this hypothesis. In a trust game experiment (where individuals play in both roles), they show that people who are untrustworthy as receivers tend to underestimate average trustworthiness in the subjects' pool, and this considerably lowers their earnings when acting as sender. That is essentially what we have in mind: People who trust too little typically tend to underinvest. For instance, a youngster who believes that the world is an unjust place dominated by opportunists will be less inclined to invest in education (since he thinks that his investment is unlikely to pay off). Similarly, why should an employee exert any

[25] Evidence includes Costa Gomez et al. (2010), Charness et al. (2011), Sapienza et al. (2013), Butler et al. (forthcoming) in the trust game, Ellingsen et al. (2010) in both trust and dictator game, Gächter et al. (2010) in a sequential voluntary contribution game, Blanco et al. (2009, 2011) in a sequential prisoner's dilemma. Moreover, the consensus effect survives even after subjects are exposed to substantial learning opportunities – as shown e.g. in Butler et al. (forthcoming).

effort beyond the legally binding minimum if he thinks that, when the time comes, the employer will promote his own buddies anyway? The notion that excessive mistrust can have important economic consequences is backed by empirical evidence obtained with survey data. Butler et al. (2014) document a hump-shaped relationship between trust and earnings. People who trust too little (or too much) face worse economic outcomes.[26] They also show that the economic loss connected with trusting too little can reach a similar order of magnitude as the income premium associated with obtaining a college degree. Overall, the data indicate that, while being more trusting increases the likelihood of being cheated, the gains generated by trusting others when trust is honored are enough to compensate for this.

In sum, trusting too little can be economically damaging. The evidence on the consensus effect implies that excessive mistrust is more likely to arise in individuals who are themselves unscrupulous. However, we have argued that there is a limit to how far this can go. If trusting/investing becomes too much of a "no brainer", this is likely to be picked up even by opportunists, who will accordingly start to invest more and thus catch up with the rest. We have shown that this mechanism may potentially account for another stylized fact, namely the observed heterogeneity in social preferences, without which the very notion of a consensus effect becomes meaningless.

Finally, although the results are presented within the context of a trust game, we believe that our insights are more general and may apply to a rich set of sequential social dilemmas and other regarding preferences. In particular, it is conceivable that our key result (a stable mix of opportunistic and non-opportunistic individuals) would carry on in the case of an ultimatum game where individuals may be either opportunists or reciprocal. A similar conjecture can be made for the sequential public good game, the control game, and the gift exchange game. Future work should be devoted to clarifying these conjectures.

[26]By exploiting cross-country heterogeneity in trustworthiness, they are also able to show that the relationship between trust and earnings cannot be explained by reverse causality. Indeed, in countries with lower average trustworthiness, income peaks at lower levels of trust.

8 Appendix

8.1 Proof of Lemma 1

Denote with $h(\pi|x,\tau)$ the posterior distribution of π given *both* x and the principal's type $\tau_P = U, O$. For a type U principal,

$$h(\pi|x,\tau_P = U) = \pi \frac{g(x|\pi)f(\pi)}{\int_{z\in\mathcal{P}} z g(x|z) dF(z)} = \frac{\pi \tilde{g}(\pi|x)}{E(\pi|x)}, \tag{14}$$

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

$$h(\pi|x,\tau_P = O) = (1-\pi) \frac{g(x|\pi)f(\pi)}{\int_{z\in\mathcal{P}} (1-z) g(x|z) dF(z)} = \frac{(1-\pi)\tilde{g}(\pi|x)}{1-E(\pi|x)}, \tag{15}$$

The last two expressions show that the principal's beliefs about π depend on her own type. Denote with $\tilde{G}$ the cumulative distribution associated with $\tilde{g}$, with τ_A the agent's type, and with $b(x,\tau_P) \equiv \Pr(\tau_A = U|x,\tau_P)$ the probability assessment that the agent is of type U made by a type τ_P principal. A type U principal believes that the agent is a type U with probability

$$b(x,U) = \frac{\int_{\pi\in\mathcal{P}} \pi^2 d\tilde{G}(\pi|x)}{E(\pi|x)} = E(\pi|x) + \frac{Var(\pi|x)}{E(\pi|x)}. \tag{16}$$

The same probability for a type O principal is

$$b(x,O) = \frac{\int_{\pi\in\mathcal{P}} \pi(1-\pi) d\tilde{G}(\pi|x)}{1-E(\pi|x)} = E(\pi|x) - \frac{Var(\pi|x)}{1-E(\pi|x)}. \tag{17}$$

□

8.2 Proof of Proposition 1

The expected utility w from participation of a type $\tau_P = U, O$ principal with objective evidence x is

$$E(w|x,\tau_P = O) = b(x,O)[\theta + \alpha] - \alpha \tag{18}$$

$$E(w|x,\tau_P = U) = b(x,U)[\theta + \alpha - \rho] + 1 + \rho - \alpha. \tag{19}$$

The difference is thus

$$E(w|x,\tau_P = U) - E(w|x,\tau_P = O) = [b(x,U) - b(x,O)](\theta + \alpha) + [1 - b(x,U)]\rho + 1. \tag{20}$$

Using (1), this reduces to (2). All terms are clearly positive. □

8.3 Stable populations with enforcement

In this section we provide proofs for a slightly more general statement of Propositions 2 and 3 of Section 2.4. In particular, the results are proved for all $\phi \geq 0$ for the case of deterrence. (The case of retribution is a straightforward extension and is thus omitted). Propositions 2 and 3 can accordingly be seen as special cases (for $\phi = 0$) of Propositions 2B and 3B proved below. We also formally provide conditions under which more deterrence may crowd in or crowd out unselfish preferences.

As in Section 4, let $X_\phi^\tau \subseteq X$ denote the set of realizations of x for which type τ chooses to participate, i.e.

$$X_\phi^\tau = \{x \in X : b(x,\tau) \geq R(\tau;\phi)\}, \tag{21}$$

where $R(\tau;\phi)$ is given by (12). Relative fitness is now given by

$$2[V_U(\pi) - V_O(\pi)] = \\ \left(\mathcal{G}(X_\phi^U,\pi) - \mathcal{G}(X_\phi^O,\pi)\right)(\theta+\alpha)(1-\phi)\left(\pi\frac{\theta+\alpha-\rho}{\theta+\alpha} - R(O;\phi)\right) - \mathcal{G}(X_\phi^O,\pi)\rho(1-\phi). \tag{22}$$

As expression (3) shows, the actual values of the posterior $b(x,\tau)$ are pinned down by the exogenously given prior. To avoid complicated expressions, we will thus directly state all conditions in terms of posterior beliefs. Condition (7) generalizes into

$$b(1,U) \geq R(U;\phi). \tag{23}$$

Proposition 2 can be restated as

Proposition 2B. *When objective evidence is coarse $\pi = 0$ is asymptotically stable iff (23) holds.*

Proof. Consider a share $\epsilon > 0$ sufficiently small of type U. A fraction $1-\epsilon$ of the population observes $x = 0$ while a fraction ϵ observes $x = 1$. Assume then $b(1,U) \geq R(U;\phi)$ so that all type U observing the high signal participate. Since $b(0,O) < b(1,O) = b(0,U) < b(1,U)$ and $R(U;\phi) < R(O;\phi)$, the only relevant cases are those summarized in the following table,

In case 1, the RHS of (22) reduces to $-\rho(1-\phi) < 0$ for all π, so that $\pi = 0$ is stable. In case 2, we have

$$2[V_U(\epsilon) - V_O(\epsilon)] = (1-\epsilon)(\theta+\alpha)(1-\phi)\left(\epsilon\frac{\theta+\alpha-\rho}{\theta+\alpha} - R(O;\phi)\right) - \epsilon\rho(1-\phi), \tag{24}$$

Case		X_ϕ^U	X_ϕ^O
1	$R(O;\phi) \leq b(0,O)$	$\{0,1\}$	$\{0,1\}$
2	$R(O;\phi) \in (b(0,O), b(1,O)]$	$\{0,1\}$	$\{1\}$
3	$R(O;\phi) > b(1,O) \ \wedge \ R(U;\phi) \leq b(1,O)$	$\{0,1\}$	$\varnothing$
4	$R(O;\phi) > b(1,O) \ \wedge \ R(U;\phi) > b(1,O)$	$\{1\}$	$\varnothing$

which is negative for ϵ sufficiently small. In case 3, relative fitness has the same sign as the first term in (24) so that the same applies. Finally, in case 4,

$$2[V_U(\epsilon) - V_O(\epsilon)] = \epsilon(\theta+\alpha)(1-\phi)\left(\epsilon\frac{\theta+\alpha-\rho}{\theta+\alpha} - R(O;\phi)\right), \tag{25}$$

which is again negative for ϵ sufficiently small. Hence, $R(U;\phi) \leq b(1,U)$ is sufficient for $\pi = 0$ to be asymptotically stable. Clearly enough, when $R(U;\phi) > b(1,U)$, no one participates. Relative fitness is thus zero. This implies that $R(U;\phi) \leq b(1,U)$ is also necessary. □

As for Proposition 3, condition (8) ($\theta + \alpha/\rho > 2$) remains unchanged. Condition (9) generalizes into

$$R(O;\phi) < \frac{\theta+\alpha-2\sqrt{\rho(\theta+\alpha-\rho)}}{\theta+\alpha}, \tag{26}$$

and (10) becomes

$$b(0,O) < R(O;\phi) \leq b(1,O). \tag{27}$$

Then,

Proposition 3B. *When objective evidence is coarse, conditions (8), (26), and (27) are necessary and sufficient for the existence of an asymptotically stable polymorphic population. Under these conditions the stable share of type U is*

$$\pi^* \equiv \frac{(\theta+\alpha)(1+R(O;\phi)) - 2\rho + \Delta_\phi}{2(\theta+\alpha-\rho)} \in (0,1),, \tag{28}$$

where $\Delta_\phi \equiv \sqrt{(\theta+\alpha)^2(1-R(O;\phi))^2 - 4\rho(\theta+\alpha-\rho)}$.

Proof. We start off by showing that $b(0,O) < R(O;\phi) \leq b(1,O)$ is necessary for a stable polymorphic population. Suppose first that $b(0,O) \geq R(O;\phi)$. This implies that all individuals participate independently of their type or the signal x they observe. Since $X^U = X^O = \{0,1\}$, relative fitness (22) is strictly negative for all $\pi \in (0,1)$. As a result,

no stable polymorphic population exists. Suppose then that $b(1,O) < R(O;\phi)$. This implies $X^O = \varnothing$ so that type O never participate. Equation (22) thus becomes

$$2[V_U(\pi) - V_O(\pi)] = (\theta+\alpha)(1-\phi)\mathcal{G}(X^U,\pi)\left[\pi\frac{\theta+\alpha-\rho}{\theta+\alpha} - R(O;\phi)\right], \tag{29}$$

where $\mathcal{G}(X^U,\pi)$ is equal to one if $X^U = \{0,1\}$, is equal to π if $X^U = \{1\}$, and is equal to zero if $X^U = \varnothing$. Inspection of (29) shows that for all π such that $V_\pi(T) - V_\pi(O) = 0$, $|_{z=\pi} d(V_z(T) - V_z(O))/dz \geq 0$, so that no polymorphic population is stable.

Given $b(0,O) < R(O;\phi) \leq b(1,O)$, $X^O = \{1\}$ and $X^U = \{0,1\}$ (since $b(1,U) \geq b(0,U) = b(1,O)$). Equation (22) thus becomes

$$2[V_U(\pi) - V_O(\pi)] = (1-\pi)(\theta+\alpha)(1-\phi)\left(\pi\frac{\theta+\alpha-\rho}{\theta+\alpha} - R(O;\phi)\right) - \pi\rho(1-\phi). \tag{30}$$

Clearly enough, any stable polymorphic population, if there is any, is a root of the RHS of (30) (although not all roots are necessarily stable populations). The next Lemma gives necessary and sufficient conditions for the existence of real roots in the (0,1) interval.

Lemma 2. *The RHS of (30) has real roots in* $(0,1)$ *if and only if* $\theta+\alpha > 2\rho$ *and* $R(O;\phi) < \frac{\theta+\alpha-2\sqrt{\rho(\theta+\alpha-\rho)}}{\theta+\alpha}$. *The roots are* π^* *(as given by 28) and* $\hat{\pi} = \pi^* - \Delta_\phi/(\theta+\alpha-\rho)$. *Both* π^* *and* $\hat{\pi}$ *lie in the interval* $(R(O;\phi),1)$.

Inspection of (30) reveals that the RHS is an increasing-decreasing function taking negative values for $\pi = 0$ and $\pi = 1$. If $R(O;\phi) \geq (\theta+\alpha-\rho)/(\theta+\alpha)$, then $V_U(\pi) - V_O(\pi) < 0$ for all $\pi \in (0,1)$ so that the RHS of (30) has no real root in $(0,1)$. However, if $R(O;\phi) < (\theta+\alpha-\rho)/(\theta+\alpha)$ and $\sqrt{(\theta+\alpha)^2(1-R(O;\phi))^2 - 4\rho(\theta+\alpha-\rho)} > 0$, then the RHS of (30) has two real roots,

$$\hat{\pi}, \pi^* = \frac{(\theta+\alpha)(1+R(O;\phi)) - 2\rho \pm \sqrt{(\theta+\alpha)^2(1-R(O;\phi))^2 - 4\rho(\theta+\alpha-\rho)}}{2(\theta+\alpha-\rho)}. \tag{31}$$

of which the largest is π^* as given by (28). It is immediate to check that $\hat{\pi} < \pi^* < 1$, and that $\pi^* > \hat{\pi} > 0$ requires $R(O;\phi) > (2\rho-\theta-\alpha)/(\theta+\alpha)$. To sum up, $R(O;\phi)$ must satisfy

$$\frac{2\rho-\theta-\alpha}{\theta+\alpha} < R(O;\phi) < \min\left\{\frac{\theta+\alpha-2\sqrt{\rho(\theta+\alpha-\rho)}}{\theta+\alpha}, \frac{\theta+\alpha-\rho}{\theta+\alpha}\right\}. \tag{32}$$

Notice that

$$\frac{2\rho-\theta-\alpha}{\theta+\alpha} < \frac{\theta+\alpha-2\sqrt{\rho(\theta+\alpha-\rho)}}{\theta+\alpha} \tag{33}$$

only if $\theta + \alpha > 2\rho$. Moreover, given a) $R(O;\phi) > 0$ (since $R(O;\phi) > b(0,O) \geq 0$ by assumption) and b) $\theta + \alpha > 2\rho$, the first inequality in (32) always holds and can be ignored. Under $\theta + \alpha > 2\rho$ we also have

$$\frac{\theta + \alpha - \rho}{\theta + \alpha} > \frac{\theta + \alpha - 2\sqrt{\rho(\theta + \alpha - \rho)}}{\theta + \alpha}. \tag{34}$$

Hence, $R(O;\phi) < \frac{\theta+\alpha-2\sqrt{\rho(\theta+\alpha-\rho)}}{\theta+\alpha}$ and $\theta + \alpha > 2\rho$, are necessary and sufficient conditions for the RHS of (30) to have two real roots in the interval $(0,1)$. Moreover $\theta + \alpha > 2\rho \Rightarrow \pi^* > \hat{\pi} > R(O;\phi)$. This proves Lemma 2.

The proof of Proposition 3B is concluded by noticing that $d(V_U(z) - V_O(z))/dz|_{z=\hat{\pi}} \geq 0$ and $d(V_U(z) - V_O(z))/dz|_{z=\pi^*} < 0$, so that $\hat{\pi}$ is unstable and π^* is stable. □

The crowding in and crowding out results immediately follow from Propositions 2B and 3B.

Corollary 1 (*crowding in*)

a) Assume $\theta + \alpha > 2\rho$ and consider two levels of enforcement ϕ' and ϕ'' with $\phi'' > \phi'$. Then, a positive fraction π^* of type U is possible under ϕ'' but not under ϕ' if

$$b(0,O) < R(O;\phi'') < \min\left\{b(1,O), \frac{\theta + \alpha - 2\sqrt{\rho(\theta + \alpha - \rho)}}{\theta + \alpha}\right\} < R(O;\phi'). \tag{35}$$

b) Consider two levels of enforcement ϕ' and ϕ'' with $\phi'' > \phi'$. If conditions (8), (26), and (27) are satisfied for both ϕ' and ϕ'', then the stable share of type U, π^*, is larger under ϕ'' than under ϕ'.

Corollary 2 (*crowding out*) *Assume $\theta > 2\rho$ and consider two levels of enforcement ϕ' and ϕ'' with $\phi'' > \phi'$. Then, a positive fraction π^* of type U is possible under ϕ' but not under ϕ'' if*

$$R(O;\phi'') < b(0,O) < R(O;\phi') < \min\left\{b(1,O), \frac{\theta + \alpha - 2\sqrt{\rho(\theta + \alpha - \rho)}}{\theta + \alpha}\right\}. \tag{36}$$

References

[1] Adriani, F., and Sonderegger, S. (2009) "Why do parents socialize their children to behave pro-Socially? An Information-Based Theory" Journal of Public Economics 93: 1119-1124.

[2] Alger, I, and Weibull, J.W. (forthcoming) "A generalization of Hamilton's rule – love others how much?" Journal of Theoretical Biology.

[3] Bar-Gill, O. and Fershtman, C. (2004) "Law and preferences" Journal of Law Economics and Organization, 20: 331-352.

[4] Bar-Gill, O. and Fershtman, C. (2005) "Public policy with endogenous preferences" Journal of Public Economic Theory, 7: 841–857.

[5] Bénabou, R. and Tirole, J. (2003) "Intrinsic and extrinsic motivation" Review of Economic Studies, 70: 489-520.

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

[7] Bisin, A., and Verdier, T., (2001) "The economics of cultural transmission and the dynamics of preferences" Journal of Economic Theory, 97: 298-319.

[8] Blanco, M., Engelmann, D., Koch, A. K. and Normann, H.-T. (2009) "Preferences and Beliefs in a Sequential Social Dilemma: A Within-Subjects Analysis," IZA Discussion Paper 4624.

[9] Blanco, M., Engelmann, D. and Normann, H.-T. (2011) "A Within-Subject Analysis of Other-Regarding Preferences," Games and Economic Behavior, 72: 321-338.

[10] 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.

[11] Butler, J. V., Giuliano, P., and Guiso, L. (forthcoming) "Trust, Values and False Consensus", International Economic Review.

[12] Butler, J. V., Giuliano, P., and Guiso, L. (2013) "Trust and Cheating", Mimeo.

[13] Butler, J. V., Giuliano, P., and Guiso, L. (2014) "The Right Amount of Trust", Mimeo.

[14] Cohen, D., and Eshel, I. (1976) "On the founder effect and the evolution of altruistic traits" Theoretical population biology, 10: 276-302.

[15] Costa-Gomes, M. A., Huck, S. and Weizsäcker, G. (2010) "Beliefs and actions in the trust game: creating instrumental variables to estimate the causal effect" IZA working paper n. 4709.

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

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

[18] Engelmann, D. and Strobel, M., (2000) "The false consensus effect disappears if representative information and monetary incentives are given," Experimental Economics (3): 241-260.

[19] Engelmann, D. and Strobel, M., (2012) "Deconstruction and Reconstruction of an Anomaly," Games and Economic Behavior, 76: 678-689.

[20] Ellingsen, T., and Johanneson M. (2008) "Pride and prejudice: the human side of incentive theory" American Economic Review, 98: 990-1008.

[21] Ellingsen, T., Johanneson M., Torsvik, G. and Tjøtta, S. (2010) "Testing guilt aversion" Games and Economic Behavior, 68: 95-107.

[22] Eshel, I., Samuelson, L., and Shaked, A. (1998) "Altruists, egoists, and hooligans in a local interaction model". American Economic Review, 88:157-179.

[23] Frank, R. (1987) "If homo economicus could choose his own utility function, would he want one with a conscience?" American Economic Review, 77: 593-604. he

[24] Frey, B.S. (1997) "A constitution for knaves crowds out civic virtues" Economic Journal, 107: 1043-1053.

[25] Frey, B.S. and Jegen, R. (2001) "Motivation crowding theory" Journal of Economic Surveys, 15: 589-611.

[26] Gamba, A. (2012) "Learning and evolution of altruistic preferences in the Centipede Game" Journal of Economic Behavior and Organization, forthcoming.

[27] Gächter, S., Nosenzo, D., Renner, E. and M. Sefton (2010) "Who makes a good leader? cooperativeness, optimism and leading-by-example" Economic Inquiry, in press.

[28] Gneezy, U. (2005) "Deception: the role of consequences" American Economic Review, 95: 384-394.

[29] Goeree, J. K. and Großer, J. (2006) "Welfare reducing polls" Economic Theory 31: 51–68.

[30] 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.

[31] 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.

[32] Huck, S., and Oechssler, J. (1999) "The indirect evolutionary approach to explaining fair allocations." Games and Economic Behavior 28: 13-24.

[33] Orbell, J., and R.M. Dawes (1991) "A 'cognitive miser' theory of cooperators' advantage" American Political Science Review. 85: 515-528.

[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., and Samuelson, L., (2010) "The evolutionary foundations of preferences" in Handbook of Social Economics. Eds. A. Bisin and M. Jackson, 221-310, North-Holland.

[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. (2004) "Information-based relative consumption effects" Econometrica. 72: 93-118.

[38] Samuelson, L. (2005) "Economic theory and experimental economics," Journal of Economic Literature, 43: 65-107.

[39] Samuelson, L., and Swinkels, J. (2006) "Information, evolution and utility" Theoretical Economics. 1:119-142.

[40] Samuelson, L., and Zhang, J., (1992) "Evolutionary stability in Asymmetric Games", Journal of Economic Theory 67: 363-391.

[41] Sapienza, P., Toldra, A., and Zingales, L., (2010). "Understanding trust" Mimeo, Kellogg School of Management.

[42] Selten, R. and Ockenfels, A. (1998) "An experimental solidarity game" Journal of Economic Behavior and Organization 34: 517-539.

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

[44] Weibull, J. W. (1997). Evolutionary game theory. MIT press.