CEDEX
CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Discussion Paper No. 2013-09

Fabrizio Adriani and
Silvia Sonderegger

October 2013

Trust, Trustworthiness and the Consensus Effect: An Evolutionary Approach

CeDEx Discussion Paper Series
ISSN 1749 - 3293

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

Fabrizio Adriani
University of Leicester

Silvia Sonderegger
University of Nottingham and CeDEx

October 25, 2013

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. Trustworthy individuals are more "optimistic" than opportunists and are accordingly less afraid to engage in market-based exchanges, where they may be vulnerable to opportunistic behavior. In some cases, the material benefits from greater market participation may compensate for the costs of being trustworthy. We use an indirect evolutionary approach to endogenize preferences for trustworthiness, showing that a polymorphic equilibrium (where both trustworthiness and opportunism coexist in the population) may be evolutionarily stable. Better institutions limiting the scope for opportunism may favor the spreading of trustworthiness (*crowding in*), but the opposite (*crowding out*) may also occur. Our analysis is consistent with experimental evidence.

Jel Codes: C73, D02, D03, D82, Z1.

Keywords: Endogenous Preferences, Trust, Consensus Effect, Institutions, Crowding Out.

*We 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.

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 recently started to pay attention to this phenomenon. A recent experiment by Sapienza et al. (2010), for instance, suggests that people playing a trust game tend to extrapolate their opponent's behavior from their own. Blanco at al. (2009, 2011) and Gächter et al. (2010) find similar evidence in experiments involving a sequential prisoner's dilemma and a sequential voluntary contribution game, respectively.[1] Experimental evidence by Ellingsen et al. (2010) shows that the tendency to extrapolate the opponent's beliefs from one's own is responsible for a significant share of what had been previously believed to be evidence of guilt-aversion.

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. 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. Indeed, Engelmann and Strobel (2000, 2012) distinguish between the *true* consensus effect – i.e., consistent with Bayesian updating – and the *false* consensus effect – which emerges even among individuals that possess the same information. Within an experimental setting, they show that the consensus effect disappears if people are told the distribution of behaviors within their interacting group. This suggests that the effect is primarily driven by information imperfections, and corroborates the true consensus effect hypothesis.

The aim of this work is to study the implications of the consensus effect for trust and trustworthiness, from a theoretical perspective. We mainly focus on the "true" consensus effect, which is consistent with Bayesian learning.[2] Most economic models assume that

[1] See also Selten and Ockenfeld (1998). Costa-Gomes et al. (2010) use an instrumental variable approach to prove that there is indeed a causality relationship between the trustor's beliefs and his actions in the trust game.

[2] Alternatively, we could have allowed for an irrational (false) consensus effect (such as, e.g., a systematic bias). As will become clear below, the main forces at work in the analysis (the selection effect and the expropriation advantage) would have remained unchanged, although, clearly, the conclusions would have been affected by the exact modelling details of the irrational consensus effect.

the distribution of types within a population is common knowledge.[3] This implies that an individual's own type provides no relevant information about it, and, thus, individual beliefs about the frequency of different types must be type-independent. However, the assumption that individuals are able to observe the distribution of types within the population is strong. If it is relaxed, then it becomes rational for individuals to use their own types to make inferences about the overall population.

A detailed description of our model can be found in Section 2. We consider a setup where individuals can be of two types: opportunistic or trustworthy. Trusting is optimal only when one's counterparty is trustworthy. However, the type of an individual is unobservable by others. Players thus choose to trust or not depending on their beliefs about the composition of the overall population. We assume that individuals have access to *objective* sources of information about the share of trustworthy, but these are noisy. As a result, a rational individual will also take into account the information conveyed by her own type (*introspection*). This captures the idea that individuals look at the way *they* would behave in a certain situation in order to make predictions about the way *their counterparty* is likely to behave in the same situation (the *consensus effect*). When facing the same objective evidence, trustworthy individuals have accordingly a higher propensity to believe that others are trustworthy, and are therefore more inclined to trust them, than the opportunists. The correlation between trust and trustworthiness implies that individuals are more likely to engage in market interactions, in which they may be vulnerable to opportunistic behavior by ruthless individuals, when they are themselves trustworthy. This generates a *selection effect*, which is at the centre of our analysis. Butler et al. (2009) provide experimental evidence indicating that the selection effect is indeed sizeable.

In Section 3, we endogenize the long-term ethical attitudes by postulating the presence of an evolutionary selection process. This could reflect genetic evolution but also cultural evolution (which may presumably operate more quickly). We show that a monomorphic population made entirely of opportunists is evolutionarily stable. However, we also characterize the conditions for a polymorphic population to be evolutionarily stable, implying that both trustworthy and opportunistic preferences may coexist in the long run. This

[3] Ellingsen and Johanneson (2008) present an exception to this. They build a model where players' beliefs about the opponent's type may be positively correlated with their own type. See also Adriani and Sonderegger (2009) for a framework where the consensus effect arises as an equilibrium of the game where parents select the values to instill in their children.

is consistent with an accepted “stylized fact” of the experimental literature, namely that there is widespread and substantial heterogeneity in preferences.

Since our setup explicitly rules out the two features that are usually conducive to the survival of unselfish preferences, namely assortative matching and observability of preferences, the long-term survival of trustworthy preferences may appear surprising at first glance. Trustworthy individuals fail to expropriate others, which puts them at a comparative disadvantage. However, we show that the selection effect also affords a potential advantage to the trustworthy, since they are more likely to engage in market interactions. Depending on the actual composition of the population, this may compensate for the cost of foregoing lucrative expropriation opportunities. This raises the question of whether the trustworthy can fully displace the opportunists. While this is possible in particular cases, our model highlights the presence of countervailing forces setting a natural upper bound to the share of trustworthy in the population. As the trustworthy spread, all individuals (including the opportunists) become more likely to observe objective evidence suggesting that trusting is indeed optimal. The opportunists become accordingly more willing to trust, so that the selection effect is weakened. In other words, the very prevalence of the trustworthy undermines their evolutionary advantage. This results in a stable polymorphic population where trustworthy individuals do materially as well as opportunistic ones. Butler et al. (2009) provide empirical evidence that supports this hypothesis. They 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.[4]

We also analyze the interaction between ethical attitudes and institutions aimed at limiting the scope of opportunistic behavior. In Section 4, we show that, although better institutions do in some cases favor the spreading of trustworthy ethical attitudes in the long-run, this is not always the case. The intuitive reason is that institutions may “crowd out” intrinsic trustworthiness by weakening the selection effect. Our results indicate that the long-run effects of an improvement of the institutional environment may be very different from the short run effects. In the short run the distribution of preferences (ethical

[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, their simple framework does not consider the evolutionary implications of this advantage. By contrast, in our model the fraction of pro-social individuals is determined endogenously.

attitudes) is fixed, while 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" when preferences are endogeneous.[5] This mechanism differs from the motivational crowding out identified in the literature.[6]

Section 5 addresses a number of extensions to our baseline framework. In Section 5.1 we show that our mechanism can be applied to rationalize the survival of various types of social preferences, like altruism, reciprocal altruism, and homophily. In Section 5.2, we study what happens when individuals are able to obtain progressively finer and more accurate information about the composition of the population. So long as this information is less than perfect, a polymorphic equilibrium continues to exist, although the set of suitable parameters for this to be the case grows progressively smaller. Finally, Section 6 concludes by describing what we believe are our key contributions to the existing literature, and by discussing future work.

2 Exogenous preferences

2.1 The one-shot game

Principals 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, which gives her a material welfare equal to $\alpha > 0$. If the principal chooses to participate (p) and the agent does not cheat (nc), the principal will

[5] The existence of crowding out effects is well documented in the empirical literature. Fehr and Gächter (2002) present experimental evidence that incentive contracts may undermine voluntary cooperation. Also within an experiment, Bohnet et al. (2001) show that greater contractibility may crowd out trustworthiness. Huck (1998) and Bar-Gill and Fershtman (2004 and 2005) build models where, as in ours, preferences are derived endogenously and may be affected by the institutional environment. However, the mechanisms at work are very different from ours.

[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.

obtain $\theta > \alpha$. In contrast, if the agent cheats (c), the principal obtains zero. Hence, in this latter case, the principal would have been better off not participating in the exchange at all.[7] 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.

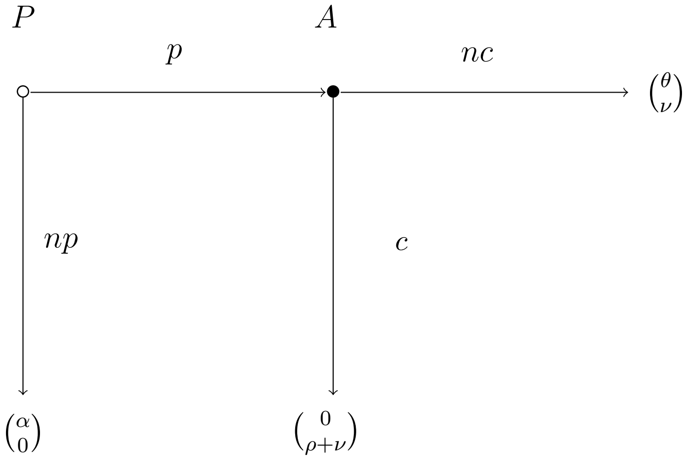

Figure 1: Material payoff game

Agents If the principal chooses not to participate, the agent receives a material payoff equal to zero. When the principal participates, an agent obtains $\nu \geq 0$ if he does not cheat and $\nu + \rho$, $\rho > 0$, if he cheats. An agent's material welfare is thus maximized by cheating whenever trusted. We assume $\rho < \theta$ – i.e. engaging in opportunistic behavior is inefficient. This assumption is necessary for the long-term survival of preferences for trustworthy behavior. 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.

Preference Traits We allow for individuals who may have preferences that do not coincide with their selfish material payoff. More specifically, we assume that individuals may be of two types: opportunistic (O) and trustworthy (T). Type O individuals only care about selfish material welfare. In contrast with type O, type T individuals have other

[7] 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. Our results would however carry through if we were to use their setup. 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 6).

regarding preferences. We show in Section 5.1 that our results are compatible with a variety of other regarding motives including altruism and reciprocity. However, in order to isolate the core mechanism of our theory, we abstract here from the precise nature of type T's other regarding preferences. We simply assume that type T do not cheat when acting as agents.[8] For the same reason, when playing as principals, both types are assumed to have the same objective function, i.e. they maximize their own expected material payoff. We stress that this is purely done in order to rule out confounding effects that would only strengthen the mechanism at work in the model (see Section 5.1).

2.2 Information and beliefs

Following Dawes (1989), we now 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 common knowledge that the population contains both type T and type O individuals, however the precise share of is not known with certainty. This is a crucial assumption since it ensures a role for introspection. 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 trustworthy.

We denote with π the share of type T 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, the principal has two pieces of relevant information. First, she observes a noisy signal $x \in X$ about π, which captures the information that the principal is able to collect about the composition of the general population. 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. Conditional on π, x has density $g(x|\pi)$ and cumulative

[8] 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 the trustworthy. Intuitively, a trustworthy individual may be better equipped to figure out whether stakes are so high that even the trustworthy may be tempted to cheat.

$G(x|\pi)$, 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.[9] We assume $Var(\pi|x) > 0$ for all $x \in X$, so that there is no realization of x which fully reveals π.

In addition to the signal x, another piece of information that is available to a principal is her own type, τ. Summarizing, the timing of the game is as follows,

1. Nature randomly assigns each individual in the population to a role (P/A) and then draws a principal and an agent to play the game.
2. Players observe their own type $\tau \in \{T, O\}$. 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 consensus effect

The mechanism whereby Bayesian learning generates a consensus effect may be illustrated through a simple example. Suppose that it is common knowledge that the population is perfectly homogenous, i.e. it is either totally composed of type T or of type O with equal probabilities. Formally, the common prior on π is that $\pi = 1$ with probability 1/2, $\pi = 0$ with probability 1/2, and $\pi \in (0, 1)$ with probability zero. Suppose now that an individual is able to observe whether she is of type T or O. Given her prior, this would allow her to perfectly predict the type of any other individual drawn from the same population. This shows that, despite the prior being the same for all, posterior beliefs are type-dependent.

We now generalize this intuition by allowing both for a generic prior (including possibly uninformative priors) and for objective evidence (i.e. the signal x). Denote with $b(x, \tau_P) \equiv \Pr(\tau_A = T|x, \tau_P)$ the probability assessment that the agent is of type T made by a type τ_p principal who observes a signal realization x.

[9] From Bayes' rule,

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

and

$$Var(\pi|x) = \int_{\pi \in \mathcal{P}} (\pi - E(\pi|x))^2 \frac{g(x|\pi)dF(\pi)}{\int_{z \in \mathcal{P}} g(x|z)dF(z)}. \tag{2}$$

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

$$b(x,T) = E(\pi|x) + \frac{Var(\pi|x)}{E(\pi|x)} > b(x,O) = E(\pi|x) - \frac{Var(\pi|x)}{1 - E(\pi|x)},\ \forall x \in X \tag{3}$$

Proof. See Appendix.

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

2.4 The selection effect

Consider now equilibrium play. A strategy for the principal maps her information $\{T,O\} \times X$ into a distribution over (p, np). A strategy for the agent maps $\{T,O\} \times (p, np)$ 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 while, by assumption, type T agents play nc. By cheating, type O agents maximize their material welfare, whereas type T "leave money on the table". We refer to this feature of the model as the opportunists' *expropriation advantage.*

Consider now principals. The consensus effect immediately implies that, 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.*

Corollary 1. *(The selection effect) Given the same objective evidence, a trustworthy principal has a higher propensity to participate than an opportunistic principal, i.e. a trustworthy participates whenever an opportunistic would participate, but the reverse is not generally true.*

Proof. A type τ_P principal with objective evidence x chooses to participate if the expected net payoff w from participating is positive[10], i.e.

$$E(w|x,\tau_P) = b(x,\tau_P)\theta - \alpha \geq 0 \tag{4}$$

[10] We assume for simplicity that the principal chooses to participate when indifferent. This has no consequence for our results.

Expression (3) shows that the difference in expected net payoffs between a type T and a type O principals can be written as

$$E(w|x, \tau_P = T) - E(w|x, \tau_P = O) = \theta \frac{Var(\pi|x)}{E(\pi|x)(1 - E(\pi|x))} > 0 \tag{5}$$

which is always positive. □

From expression (5), notice that the differences in the two types' expected material payoff from participation is proportional to $Var(\pi|x)/\left[E(\pi|x)(1 - E(\pi|x))\right]$. This ratio 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.[11] Overall, a large value for the ratio indicates that introspection is a relatively accurate signal of π. Hence, the beliefs about the agent's trustworthiness held by a principal of type T will be much more optimistic than those of a type O. By contrast, a small value for the ratio implies that beliefs will not depend much on the principal's type.

2.5 Coarse information

We now provide a detailed illustration of the consensus and the selection effects 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. $X = \{1, 0\}$. Depending on their information, individuals may be of four types ($\{T, O\} \times \{0, 1\}$): trustworthy who observed a high or low realization of x and opportunists with a high or low realization. Let $g(x = 1|\pi)$ be the probability of receiving the high signal. Clearly, when a fraction π of individuals are of type T, $g(x = 1|\pi) = \pi$. Symmetrically, the probability of receiving the low signal, $g(x = 0|\pi)$, is $1 - \pi$. Given the information structure, the signals x and τ

[11] 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,

$$Var(\boldsymbol{\tau} \mid x) = E\left(Var(\boldsymbol{\tau} \mid \pi) \mid x\right) + Var(E(\boldsymbol{\tau} \mid \pi) \mid x) \tag{6}$$

Straightforward calculations show that the first term in (6) can be written as

$$E\left(Var(\boldsymbol{\tau} \mid \pi) \mid x\right) = E(\pi|x)(1 - E(\pi|x)) - Var(\pi|x). \tag{7}$$

Moreover, since $E(\boldsymbol{\tau} \mid \pi) = \pi$, the second term in (6) is equal to $Var(\pi|x)$. Hence, $E(\pi|x)[1 - E(\pi|x)]$ is equal to $Var(\boldsymbol{\tau} \mid x)$.

are identically and independently distributed conditional on π. Simple calculations show that the conditional expectation of π given either signal (x or τ) is

$$E(\pi|x=1)=E(\pi|\tau=T)=\frac{\Pi_2}{\Pi_1},\ E(\pi|x=0)=E(\pi|\tau=O)=\frac{\Pi_1-\Pi_2}{1-\Pi_1} \tag{8}$$

where $\Pi_n \equiv E(\pi^n)$ is the n-th moment about the origin of the prior $F(\pi)$. The conditional variances are

$$\begin{aligned} Var(\pi|x &= 1)=Var(\pi|\tau=T)=\frac{\Pi_1\Pi_3-\Pi_2^2}{\Pi_1^2} \\ Var(\pi|x &= 0)=Var(\pi|\tau=O)=\frac{\Pi_2(1-\Pi_2)-\Pi_3(1-\Pi_1)+\Pi_1\Pi_2-\Pi_1^2}{(1-\Pi_1)^2}. \end{aligned} \tag{9}$$

Lemma 1 thus implies

$$b(1,T)=\frac{\Pi_3}{\Pi_2}>b(1,O)=b(0,T)=\frac{\Pi_2-\Pi_3}{\Pi_1-\Pi_2}>b(0,O)=\frac{\Pi_1-2\Pi_2+\Pi_3}{1-2\Pi_1+\Pi_2}. \tag{10}$$

In order to understand these results, notice that the strict inequalities in (10) follow from $Var(\pi|x)>0$. In particular, $Var(\pi|x=1)>0$ implies $b(1,T)>b(1,O)$ and $Var(\pi|x=0)>0$ ensures $b(0,T)>b(0,O)$.

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 $\Pi_3=1/4$. In terms of posterior beliefs, this would imply the situation described in Figure 2.

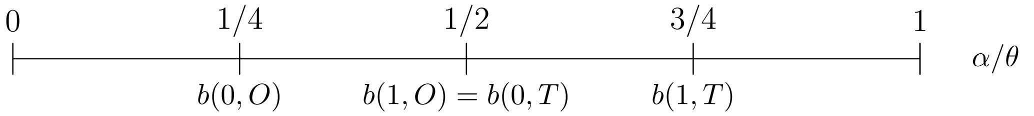

Figure 2: Uniform prior and binary signal.

A type τ principal would choose to participate only if $b(x,\tau)\geq\alpha/\theta$. If the ratio α/θ is below 1/4, then all individuals acting as principals will participate. If it is in the interval $(1/4,1/2]$, then all but the opportunists observing $x=0$ would participate. Symmetrically, if it is in the interval $(1/2,3/4]$, then only the trustworthy observing $x=1$ would participate. Finally, if α/θ is above 3/4, no one would participate. Notice that, except for the extreme cases of full participation or no participation, the share of principals who participate in equilibrium depends on the actual share of trustworthy π. This happens for two reasons. First, type T are ceteris paribus more willing to participate. Second, the share of principals (of both types) observing the high signal realization depends on π.

3 Endogenous preferences

In this section, we endogenize the share of trustworthy and analyze the long run equilibria of the model.

At present, there is no universally accepted model of cultural evolution. According to 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. Since we are interested in characterizing how evolutionary forces operate on *preferences* (rather than behavior), we follow Dekel et al. (2007) in adopting the indirect evolutionary approach pioneered by Güth and Yaari (1992).[12] That is, we assume that individuals are not pre-programmed to adopt a certain behavior. Rather, individuals (who are endowed with opportunistic or trustworthy preferences) select their behavior rationally given their preferences and the beliefs associated (via introspection) with their preferences.

We build our evolutionary analysis under the premise that an individual's decision to trust or not depends on what she has learned (either from internal or external sources). Our model does not give a direct account for why this happens. We take it for granted and rule out situations whereby learning is irrelevant to the behavior of individuals. This may happen for instance if

i) Evolution could endow individuals with priors that are perfectly tailored to π.

ii) Evolution could "hardwire" trusting behavior.

We discuss the plausibility of these cases in Section 5.3.

3.1 Extended model

We start by adding some structure to the simple model presented in the previous section. In every period, half of the population is randomly assigned to the role of principal and the remaining half to the role of agent. Principals and agents are then matched into pairs to play the one-shot game described in the previous section.

[12] 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).

Our assumptions imply that adaptation works at the *population level* rather than at the *role level*. They thus apply to cases where the probability of ending up in a certain role does not depend much on whether one's progenitor played in the same role.[13]

To simplify the notation, we normalize the payoffs of the simplified trust game by multiplying them by a factor of 2. This allows us to omit the probability that an individual is allocated in either role (1/2) when presenting ex-ante expected payoffs.

The information structure is the same as in the previous section. The description of the extended model is completed by assuming that all individuals are born with a common non-degenerate prior which is fixed. The fact that the in-built prior does not change with the actual value of π is mostly a modelling choice. 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 signal x depends on the actual value of π, the distribution of (relevant) prior beliefs in the population depends on the actual share of trustworthy individuals in the population.[14] However, for clarity of exposition, we will refer to the in-built prior simply as "the prior".

3.2 Relative fitness

Consistent with the evolutionary literature, we will interpret material payoffs as "fitness" in the rest of the paper. Let $X^\tau \subseteq X$ denote the set of realizations of x for which type τ chooses to participate, i.e.

$$X^\tau = \{x \in X : b(x, \tau) \geq \alpha/\theta\}. \tag{11}$$

An individual i of type τ observing $x_i \in X^\tau$ will participate and obtain θ with probability π and zero otherwise. The same individual observing $x_i \notin X^\tau$ will choose not to

[13] Hence, this assumption limits the scope of our results to societies where social mobility is sufficiently high. 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).

[14] We could equally assume that individuals are born with a non-degenerate prior f_π that is centered around the true value π. However, from a modelling viewpoint, this solution is far less elegant. For instance, what would the shape of f_π be when the true value of π approaches the upper or lower bounds of the $[0, 1]$ interval? Moreover, this approach raises theoretical issues since individuals need to know π (a parameter of the function f_π) in order to estimate the value of π. Our modelling approach bypasses these problems since, through the signal x, the "relevant prior" distribution only depends on π through the signal generating process.

participate and will obtain α for sure. As agents, type T individuals obtain a material payoff equal to ν whenever trusted, so that their total average fitness is

$$V_\pi(T) \equiv \mathcal{G}(X^T,\pi)[\pi\theta - \alpha] + \alpha + \nu\left\{\pi\mathcal{G}(X^T,\pi) + (1-\pi)\mathcal{G}(X^O,\pi)\right\}, \qquad (12)$$

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

$$V_\pi(O) \equiv \mathcal{G}(X^O,\pi)[\pi\theta - \alpha] + \alpha + (\nu+\rho)\left\{\pi\mathcal{G}(X^T,\pi) + (1-\pi)\mathcal{G}(X^O,\pi)\right\}. \qquad (13)$$

We can then write the difference between type T's and type O's average fitness as a function of the *actual* share of trustworthy individuals in the population,

$$V_\pi(T) - V_\pi(O) = \left(\mathcal{G}(X^T,\pi) - \mathcal{G}(X^O,\pi)\right)(\pi(\theta-\rho)-\alpha) - \mathcal{G}(X^O,\pi)\rho. \qquad (14)$$

In the jargon of evolutionary biology, $V_\pi(T) - V_\pi(O)$ is the *relative fitness* of type T given π. For any share of trustworthy individuals, a positive value for relative fitness causes the fraction of type T to increase, while a negative value causes it to shrink. As is common in the evolutionary literature, we focus on populations that are asymptotically stable in the replicator dynamic.

Definition 1. *A population π is asymptotically stable if either 1) (Monomorphic population) $\pi = 1$ ($\pi = 0$) and, for $\epsilon > 0$ sufficiently small, $V_{1-\epsilon}(T) > V_{1-\epsilon}(O)$ ($V_\epsilon(O) > V_\epsilon(T)$), or 2) (Polymorphic population) $V_\pi(T) = V_\pi(O)$ and $d(V_z(T) - V_z(O))/dz|_{z=\pi} < 0$.*

A monomorphism occurs if the population is entirely composed by individuals with one trait. Asymptotic stability requires that rare mutants obtain lower fitness than the incumbent trait. A polymorphism arises when the two traits coexist ($\pi \in (0,1)$). For a polymorphic population π to be stable, both traits must have the same fitness and, after a small shock, the share of type T must revert to π. Under the standard replicator dynamic, this involves a negative slope for relative fitness around π.

If participation decisions are identical – namely, $X^T = X^O$ – then the first term in (14) 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 Lemma 1 that trustworthy individuals have a higher propensity to participate, i.e. $X^O \subseteq X^T$. If $\pi(\theta - \rho) > \alpha$, then the first term of (14) is (weakly) positive due to the selection effect. The sign of (14) is thus ambiguous and may change depending on π. This suggests that there may exist stable populations with a positive fraction of type T. We now provide a full characterization of the stable populations for the case of coarse information.

3.3 Stable populations with coarse information

We start from the simple case of a binary signal $x \in X = \{0, 1\}$ introduced in Section 2.5. This allows for a full characterization of the equilibrium. In section 5.2 we discuss 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 7.2 of the Appendix, and are therefore omitted.

The main variable of interest is the equilibrium share of trustworthy π. The first result we present is quite straightforward.

Proposition 1. *When objective evidence is coarse, $\pi = 0$ is asymptotically stable iff $b(1, T) \geq \alpha/\theta$.*

If the population is entirely composed of opportunists, participating is suboptimal. Since greater propensity to participate is the only potential advantage of the trustworthy over the opportunists, a population of opportunists cannot be invaded. The condition $b(1, T) \geq \alpha/\theta$ ensures that type T would be willing to participate when observing the high signal. Since participation is suboptimal when type T are rare, a rare type T mutant has strictly lower average fitness than a type O. If $b(1, T) < \alpha/\theta$, no individual ever participates and both traits have identical fitness. In this case, $\pi = 0$ is said to be *neutrally stable.*

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

$$\frac{\theta}{\rho} > 2, \tag{15}$$

$$\alpha < \theta - 2\sqrt{\rho(\theta - \rho)}, \tag{16}$$

and suppose that beliefs satisfy

$$b(0, O) < \frac{\alpha}{\theta} \leq b(1, O). \tag{17}$$

Proposition 2. *When objective evidence is coarse, conditions (15), (16), and (17) are necessary and sufficient for the existence of an asymptotically stable polymorphic population. Under these conditions the stable share of type T is*

$$\pi^* \equiv \frac{\theta + \alpha - 2\rho + \Delta}{2(\theta - \rho)} \in (0, 1), \tag{18}$$

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

Proposition 2 shows that, in the long run, different preferences may persist in the population. Heterogeneity of behavior seems to be one of the robust findings of the experimental literature (see, e.g., Samuelson 2005). However, this regularity has received little attention by theorists.

In order to gather intuition on how the polymorphic equilibrium may arise, note that condition (17) ensures that type O principals participate only when they observe $x = 1$, while type T, 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 trustworthy 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. Through the selection effect, this benefits the trustworthy (who are more likely to participate) more than the opportunists, thereby increasing relative fitness. The third effect is a purely informational effect. An increase in π makes the high signal more common, thus increasing participation by the opportunists. This weakens the selection effect. If π is large, so that participation is optimal, relative fitness is accordingly reduced.[15]

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 the selection effect actually hurts the trustworthy. Relative fitness is also negative above

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

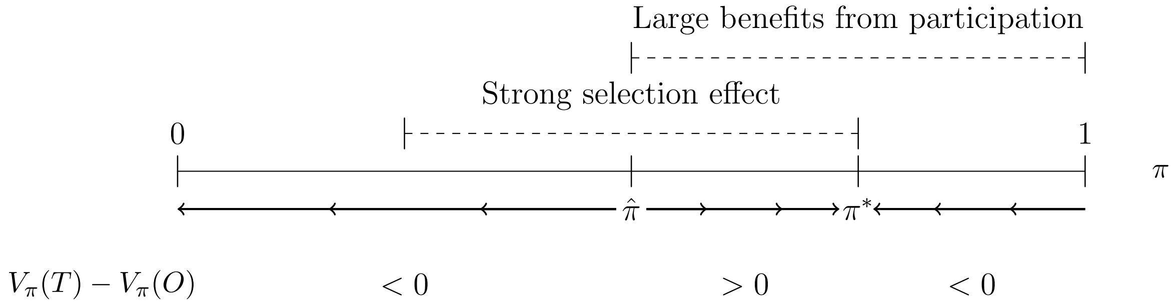

Figure 3: Stable polymorphic population.

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 selection effect has little bite. Moreover, further increases in π tend to favor the opportunists more than the trustworthy. For intermediate values of $\pi \in (\hat{\pi}, \pi^*)$, the expected benefits from participation are sufficiently large and the selection 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 T 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 (15) and (16).[16]

Consider now the intuition for why α/θ has to belong to the interval $(b(0,O), b(1,O)]$ (condition 17). If $\alpha/\theta \leq b(0,O)$, then even the opportunists who observe the low signal $x = 0$ are willing to participate. As a result, the selection effect has no bite. If $\alpha/\theta > b(1,O)$, then type O never participate. In this case, the expropriation advantage is the dominant effect for low values of π, while the selection effect dominates for sufficiently large values of π. As a result, any interior stationary point where the two effects offset each other is unstable: A small perturbation of the share of type T would lead the dynamics

[16] The requirement $\alpha < \theta - 2\sqrt{\rho(\theta-\rho)}$ in (16) can equivalently be written as $(\theta-\alpha)/\rho > 2\sqrt{(\theta-\rho)/\rho}$. The LHS captures the net (social) benefits from participation $(\theta - \alpha)$ relative to the private gain from cheating (ρ). The RHS is a concave function of the net social cost of cheating $(\theta-\rho)$ relative to the private gain from cheating. Intuitively, in our model, cheating destroys surplus both directly (given $\rho < \theta$) and indirectly (by making people more reluctant to participate). The condition suggests that a polymorphic population is more likely to emerge in environments where the direct costs of cheating are small compared to its indirect costs.

away from the stationary point.[17]

Below we provide two simple numerical examples of polymorphic equilibria.

Example A Consider the example presented in Section 2.5 with uniform prior. Assume the following parameter values: $\theta = 80$, $\alpha = 22$, $\nu = 40$, and $\rho = 10$. Given the parameters, $\alpha/\theta = 0.275$, which falls between $b(0, O)$ and $b(1, O)$. This implies that type O participate only when observing the high signal, while type T always participate. It is immediate to check that conditions (15-16) hold, so that there is a stable polymorphic population where the share of trustworthy is $\pi^* = 0.755$, i.e. nearly 76% of the population are trustworthy, with a bit more than 24% of opportunists. Given $\pi^* = 0.755$, the average payoff in the population is 98.03.[18]

Example B An alternative possibility to a uniform prior is that people may come with in-built prior information that is consistent with the long run equilibria of the model. To model this, we may choose prior distributions that assign larger weights to values of π that are close to stable populations (i.e. $\pi = 0$ and $\pi = \pi^*$). For simplicity, consider the limiting case of a prior assigning probability $a \in (0, 1)$ to $\pi = \pi^*$, probability $1 - a$ to $\pi = 0$ and probability zero to all other values of π. Clearly enough, upon observing either $\tau = T$ or $x = 1$, a Bayesian individual will infer that $\pi = \pi^*$. Hence, $b(1, O) = b(0, T) = b(1, T) = \pi^*$. Only individuals of type O observing $x = 0$ will be in any doubt as to what the actual value of π is. Their probability assessment that their agent is trustworthy is $b(0, O) = \pi^*(1 - \pi^*)^2 a/[(1 - \pi^*)^2 a + 1 - a]$. Suppose that the payoff parameters are the same as in example A. Conditions (15) and (16) are satisfied, so that a candidate for a polymorphic equilibrium is $\pi^* = 0.755$. If a is not too large, $\alpha/\theta = 0.275$ will fall between $b(0, O)$ and $b(1, O) = \pi^*$ so that condition (17) is also satisfied. For example, a natural value for a is given by the basin of attraction of the polymorphic equilibrium which is roughly equal to 0.42 and yields $b(0, O) \simeq 0.031 < \alpha/\theta$.

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

[18] The average payoff is given by

$$[\pi^* + \pi^*(1 - \pi^*)][\theta\pi^* + \rho(1 - \pi^*) + \nu] + \alpha(1 - \pi^*)^2 \tag{19}$$

where $\pi^* + \pi^*(1 - \pi^*) \simeq 0.94$ is the share of the population who participate, $\theta\pi^* \simeq 60.43$ is the average payoff from participating, $\rho(1 - \pi^*) \simeq 2.45$ is the average payoff from cheating times the share of the population who cheat, and $\alpha(1 - \pi^*)^2 \simeq 1.32$ is the payoff from not participating times the share of the population who do not participate.

Finally, for given configurations of the parameters, there also exist stable monomorphic populations entirely composed of trustworthy. It is straightforward to verify that this happens when $b(1,O) < \alpha/\theta \leq b(1,T)$ and $\theta > \alpha + \rho$. 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 trustworthy individuals would use introspection and choose not to trust even when objective evidence suggests otherwise. In the working paper version of this manuscript, we discuss these equilibria more extensively.[19] 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.

4 The role of institutions

We now introduce institutions in our model. We assume that the institutional environment only determines the extent to which market interactions are protected from opportunistic behavior. In spite of its simplicity, this approach generates a rich set of predictions.

We assume that after the principal chooses to participate, the agent has the option to cheat only with probability $1-\phi$. With probability ϕ the agent has no choice but to behave honestly.[20] The case where $\phi = 0$ corresponds to the scenario of no institutions analyzed above. When $\phi \geq \alpha/\theta$, it is optimal to participate 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 $\phi \in [0, \alpha/\theta)$.

Given ϕ, the expected net payoff from participation for a type τ individual observing signal x is

$$E(w|x,\tau) = b(x,\tau)\theta(1-\phi) + \theta\phi - \alpha. \tag{20}$$

Hence, an individual observing the pair $\{x, \tau\}$ will be willing to participate if

$$b(x,\tau) \geq \frac{\alpha - \theta\phi}{\theta(1-\phi)} \equiv R(\phi). \tag{21}$$

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

[20] Our approach shares similarities with Tabellini (2008), where the quality of institutions is modelled by the probability of detection. In our case, however, enforcement is preventive in nature.

where $R(\phi)$ is strictly *decreasing* in ϕ and ranges between α/θ (when $\phi = 0$) and 0 (when $\phi = \alpha/\theta$). As one would expect, better institutions increase the propensity to participate of both types.

On the other hand, better institutions tend to weaken the selection effect by making participation decisions less type-dependent. The difference between the net payoff from participation expected by a type T principal and that expected by a type O principal is

$$E(w|x, \tau_P = T) - E(w|x, \tau_P = O) = (1 - \phi)\theta \frac{Var(\pi|x)}{E(\pi|x)(1 - E(\pi|x))} > 0 \qquad (22)$$

Compared to (5), a positive ϕ reduces the difference in the net payoff expected by a trustworthy and an opportunistic principal. Intuitively, as institutions become 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. This weakens the selection effect.

In order to understand what this implies for the long term distribution of preferences, we need to generalize Propositions 1 and 2 to take institutions into account. This is done in the Appendix. In this section, we provide a graphical illustration of the complex interaction between institutions and trustworthiness.

Figure 4 shows the relationships between ϕ and the equilibrium share of type T (top part), and between ϕ and overall welfare (bottom part). For parameter values, the long term survival of the trustworthy may not be possible without relatively good institutions. As shown in Figure 4, if ϕ is below a certain threshold (so that $R(\phi) > (\theta - 2\sqrt{\theta(\rho - \theta)})/\theta$), the unique asymptotically stable state is $\pi = 0$. Moreover, π^* is an increasing function of ϕ. Hence, provided that the equilibrium has a positive share of trustworthy individuals, the share of trustworthy is increasing in the quality of institutions. These remarks point to a complementarity between institutions and trustworthiness, which we refer to as *crowding in*.

However, closer inspection of Figure 4 reveals that there is also another, more subtle effect at work. When ϕ exceeds another threshold (corresponding to $R(\phi) < b(0, O)$) the unique stable population is again $\pi = 0$. More effective institutions may thus have a *crowding out* effect. Intuitively, when institutions are ineffective, only the opportunists who observe the high signal realization choose to participate (while *all* trustworthy participate). Moreover, when the share of trustworthy is π^*, participation is sufficiently profitable to ensure that the selection effect offsets the expropriation advantage (which in turn ensures that a positive share of trustworthy may indeed survive). A polymorphic equilibrium is thus possible. By contrast, with better institutions, opportunists observing the low signal also start to participate. This eliminates the selection effect, since the opportunists are as likely to participate as the trustworthy. As a result, there is no equilibrium with $\pi > 0$.

A full characterization of the crowding in and the crowding out results is provided in the Appendix. Here, we illustrate crowding out through the numerical example discussed in the previous section (Example A).[21] Suppose that ϕ increases from 0 to 0.2, while all other parameters remain unchanged. $R(\phi)$ is equal to 0.094. Since $R(\phi) < b(0, O)$, all principals choose to participate independently of their type and their objective evidence. Hence, the selection effect has no bite. Relative fitness is strictly negative for all π, so that the only stable population is entirely composed of opportunists. The average payoff is now 64, compared with 98.03 under $\phi = 0$.[22] The reason why welfare decreases is that, by inducing worse ethical attitudes, better institutions end up encouraging more cheating.

Discussion 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, the consensus effect does not play any role. Rather, the result emerges because, as institutions become more effective, principals become willing to trust even those agents about whom they have unfavorable information. Hence, their rationale relies on the fact that people possess individual-specific information about their counterparties' types. This paper provides an alternative explanation and extends crowding out to interactions among strangers.

An implication of our theory 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. In particular, policies that benefit opportunists more than trustworthy may induce opportunism in the long run. In our framework, better institutions may end up doing

[21] It is easy to see that example B could also be used.

[22] Since all participate, the average payoff is now given by

$$\theta\phi + \rho(1 - \phi) + \nu \tag{23}$$

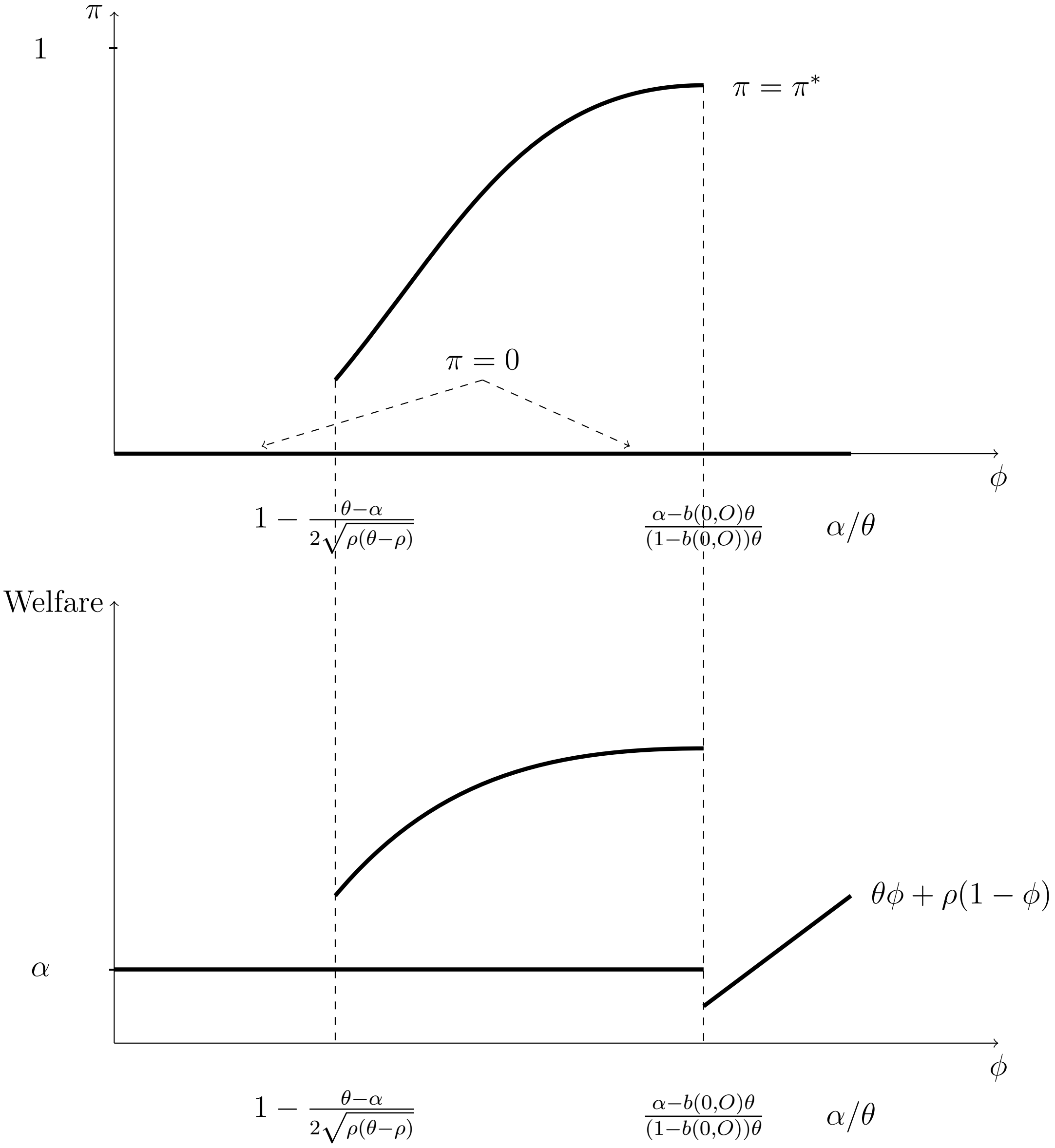

Figure 4: Share of type T and average welfare against quality of institutions. On the x-axis, the point $\phi = 1 - (\theta - \alpha)/2\sqrt{\rho(\theta - \rho)}$ corresponds to $R(\phi) = 1 - 2\sqrt{\theta(\rho - \theta)}/\theta$, the point $\phi = (\alpha - b(0, O)\theta)/(1 - b(0, O))\theta$ corresponds to $R(\phi) = b(0, O)$, the point $\phi = \alpha/\theta$ corresponds to $R(\phi) = 0$.

just that, by encouraging greater participation by opportunists.

More generally, a policy that creates a Pareto improvement in the short run may not be necessarily desirable in the long run. From a long term perspective, comparing *relative* gains is important. The key question then becomes “Who benefits more from the policy?” We believe that this is a lesson that applies beyond the specifics of the model in hand.

5 Robustness and extensions

5.1 Alternative models of social preferences: altruism, reciprocal altruism, and homophily

The assumption that type T only seek to maximize their own material welfare when acting as principals is not very plausible. An individual moved by other regarding motives would take into account the fact that his decision to trust or not affects the welfare of the agent. We accordingly extend the model to accommodate alternative preference traits for type T individuals. The main theme of this section is that explicitly incorporating into the model other regarding motives can only strengthen the selection effect, with no qualitative change in the results discussed in the previous sections. We restrict attention to the case of coarse information in order to assess the robustness of the polymorphism identified in Proposition 2.

Altruism Suppose that trait T now corresponds to a perfectly altruistic trait. Given $\theta > \rho$, an altruistic agent would always behave honestly. An altruistic principal would trust whenever

$$b(x,T)(\theta+\nu)+(1-b(x,T))(\rho+\nu)\geq\alpha\Leftrightarrow b(x,T)\geq\frac{\alpha-\nu-\rho}{\theta-\rho}\tag{24}$$

A type O principal will trust if $b(x,O)\geq\alpha/\theta$. In contrast to the previous sections, different types of principals may take different participation decisions even if they have identical beliefs. This reflects the fact that type T principals internalize their agent’s welfare, while type O principals do not.

Notice that the evolutionary dynamics only depend on realized material payoffs in the one-shot game. As a result, for Proposition 2 to apply, we only need to show that, in terms of behavior in the one-shot game, the case of altruistic preferences is observationally equivalent to the case considered in the previous sections. Suppose then that condition

(17) holds, so that type O principals participate only when observing the high realization of x. Notice that, given (17),

$$b(0,T) = b(1,O) \geq \frac{\alpha}{\theta} > \frac{\alpha - \nu - \rho}{\theta - \rho}. \tag{25}$$

This implies that type T principals observing the low signal participate. Given $b(1,T) \geq b(0,T)$, type T principals observing the high signal also participate. Equilibrium behavior in the one-shot game is thus unchanged relative to Proposition 2. The result of a polymorphic population in Proposition 2 then trivially extends to the case where the trait T represents altruism.

Reciprocal altruism Suppose now that type T individuals are reciprocal altruists. They like to behave "nicely" only when they think that their counterparty is a "nice person". A possible way to model this would be to follow Rabin (1993) and Dufwenberg and Kirchsteiger (2004) and use psychological game theory (Geanakoplos et al., 1989). This approach can however become relatively complicated in dynamic games of incomplete information. For the purposes of this discussion, we thus opt for a simpler framework, which is similar in spirit to Ellingsen and Johannesson (2008).

We start off with the case where a type T is perfectly altruistic as principal and, as an agent, reciprocates by behaving honestly only if he thinks that the principal is altruistic. Provided that type T agents reciprocate in equilibrium, the analysis of the equilibrium behavior of the principal is the same as for the case of perfect altruism. However, whether a type T agent reciprocates now depends on his beliefs about the type of the principal. Notice that a principal's decision to participate conveys information to the agent about the principal's type. The game is thus a signaling game. Assume for simplicity that the agent has no other information source than his own type and the principal's decision of whether to participate. Let μ denote a type T agent's probability assessment that his counterparty is of type T upon being matched with a principal who chooses to participate. A type T agent will reciprocate by behaving honestly whenever

$$\mu(\theta + \nu) + (1 - \mu)\nu \geq \nu + \rho \tag{26}$$

or, more simply, $\mu\theta \geq \rho$. We ignore sequential equilibria where participation never occurs. One can then verify that if

$$\mu^* \equiv \frac{\Pi_2}{2\Pi_2 - \Pi_3} \geq \rho/\theta, \tag{27}$$

and $b(0,O) < \alpha/\theta \leq b(1,O)$, then any sequential equilibrium with positive participation is such that: i) type T principals always participate, type O principals participate only when observing the high signal $x = 1$, ii) type T agents reciprocate by behaving honestly whenever trusted, type O agents cheat, iii) $\mu = \mu^*$.[23] This model is thus essentially equivalent to the one with perfect altruism. The only difference is that now we have an additional restriction on ρ and θ implied by the condition $\rho/\theta \leq \mu^*$. Notice, however, that condition (15) (necessary for a polymorphic population) already requires $\rho/\theta \leq 1/2$. Since μ^* is always greater than 1/2, the additional restriction is immaterial.

Homophily We now further extend the model of reciprocal altruism seen above. Suppose that, when playing as *principals* (as well as agents), type T are not perfectly altruistic, but only care for the welfare of their counterparty if they think that he is a “nice guy”. This model is essentially equivalent to one where type T are altruistic towards people who they think are “like them” (hence the name “homophily”). Restrict attention to equilibria with positive participation and assume $\mu^* \geq \rho/\theta$, so that type T agents always reciprocate while type O agents do not. A type T principal will only care for the welfare of agents who reciprocate. Accordingly, she will participate if

$$b(x,T)(\theta + \nu) \geq \alpha \Leftrightarrow b(x,T) \geq \frac{\alpha}{\theta + \nu} \tag{29}$$

The threshold on $b(x,T)$ necessary to induce a type T to trust is now higher than in the case of perfect altruism (but still lower than in the baseline model). This reflects the fact that a type T principal now cares only about type T agents' welfare. On the other hand, the threshold is still lower than that for opportunistic principals (α/θ). Hence, under the usual condition $b(0,O) < \alpha/\theta \leq b(1,O)$, we have that type O principals only participate upon observing the high signal $x = 1$ while T principals participate independently. The problem is thus equivalent to the case of reciprocal altruism, so that identical conclusions apply.

[23] Conditional on π, the probability that the principal participates is $\pi + \pi(1-\pi)$, the joint probability that the principal is of type T *and* participates is simply π. Finally, the probability that the agent is of type T is π. Hence, we have

$$\begin{aligned} \mu &= \int_{\pi\in\mathcal{P}} \frac{\Pr(\tau_P = T \text{ and } P \text{ participates}|\pi)\Pr(\tau_A = T|\pi)}{\int_{\pi\in\mathcal{P}}\Pr(P \text{ participates}|\pi)\Pr(\tau_A = T|\pi)dF} dF = \\ & \int_{\pi\in\mathcal{P}} \frac{\pi^2}{E(\pi[\pi + \pi(1-\pi)])} dF = \frac{\Pi_2}{2\Pi_2 - \Pi_3}. \end{aligned} \tag{28}$$

In summary, all preference traits considered have the same effect of increasing type T's propensity to participate relative to the baseline model while leaving type O's propensity unchanged. This can at most produce a strengthening of the selection effect, without qualitative differences relative to the analysis in the previous sections.

5.2 Finer information

As we have seen, for the consensus effect to work it is crucial that individuals have imperfect information about the composition of the population from which their counterparty is drawn. If the probability of being matched with a trustworthy can be accurately estimated, introspection plays a marginal role. While it appears realistic to assume that individuals lack precise information when dealing with strangers, the case of a binary signal discussed in the previous sections represents an extreme case of information coarseness. In this section we analyze what happens when the information individuals rely on becomes finer and more accurate.

To this purpose, suppose that the set of realizations of the signal x is given by $X = \{0, 1, ..., n\}$. Intuitively, the individual may observe the type of a number n of individuals in the pool, before choosing whether to participate or not. The number of type T individuals observed (x) thus has a binomial distribution with mean $n\pi$ and variance $n\pi(1-\pi)$. To keep things simple, we assume a uniform prior in $[0, 1]$ and Bayesian learning. It is then immediate to verify that the posterior probability that a type O assigns to the agent being trustworthy upon observing a realization x is $b(x, O) = (1 + x)/(3 + n)$.[24] Symmetrically, the same probability for a type T is $b(x, T) = (2 + x)/(3 + n)$. Assume for simplicity that $1/(3 + n) < \alpha/\theta < (2 + n)/(3 + n)$ and let $\hat{x}$, $0 < \hat{x} < n$, denote the realization of x such that

$$\frac{1+\hat{x}}{3+n} < \alpha/\theta \leq \frac{2+\hat{x}}{3+n} \tag{30}$$

Clearly enough, type O participate only when observing $x > \hat{x}$, while type T participate whenever $x \geq \hat{x}$. Hence, $X^O = \{\hat{x} + 1, \hat{x} + 2, ..., n\}$ and $X^T = \{\hat{x}, \hat{x} + 1, ..., n\}$. The share of type T who participate is thus given by $\mathcal{G}(X^T, \pi) = \sum_{x=\hat{x}}^{n} B(x, \pi)$, where $B(x, \pi) \equiv \binom{n}{x}\pi^x(1-\pi)^{n-x}$. The equivalent for type O is $\mathcal{G}(X^O, \pi) = \sum_{x=\hat{x}+1}^{n} B(x, \pi)$. From (14),

[24] Given the uniform prior, conditional on $x = 0, ..., n$, π has a $Beta(a, b)$ distribution with $a = x + 1$ and $b = n - x + 1$. Conditional on x, π has thus mean $E(\pi|x) = (x + 1)/(n + 2)$ and variance $Var(\pi|x) = (x + 1)(n - x + 1)/(n + 2)^2(n + 3)$. One can then use (3) to compute $b(x, T)$ and $b(x, O)$.

relative fitness is thus given by

$$V_\pi(O) - V_\pi(T) = B(\hat{x}, \pi) \left(\pi(\theta - \rho) - \alpha\right) - \rho \sum_{x=\hat{x}+1}^{n} B(x, \pi) \tag{31}$$

where the RHS can be rearranged as

$$B(\hat{x}, \pi) \left(\pi\theta - \alpha\right) - \rho \left[\sum_{x=\hat{x}+1}^{n} B(x, \pi) + \pi B(\hat{x}, \pi) \right]. \tag{32}$$

The first term in in (32) captures the difference in average net payoffs, when observing $\hat{x}$ favorable realizations, between trustworthy principals (who participate) and opportunistic principals (who do not). The second term reflects the expropriation advantage. As n becomes large, information becomes finer and more accurate. As a result, the share of individuals observing the "cutoff" realization $\hat{x}$, $B(\hat{x}, \pi)$, becomes smaller. The first term in (32), which essentially captures the selection effect, becomes accordingly less important. This is illustrated for selected values of the parameters in Figure 7, which plots relative fitness as a function of π for the cases $n = 4$ (thick line), $n = 8$ (medium line), and $n = 12$ (thin line). In all three cases, the figure shows that there exists a stable polymorphic population where type T are majoritarian. However, as n increases, the share of type T in the stable equilibrium decreases and relative fitness peaks for a lower value of π. Overall, this suggests that, from a purely quantitative viewpoint, more accurate information may generate worse ethical attitudes.

As we show in the appendix, for any finite $n \geq 1$ there exists an open set of values for the parameters of the model such that a stable polymorphic population exists. However, as n becomes large, the set of suitable parameters becomes smaller. A full characterization of the stable polymorphic equilibria for generic n is complicated by the fact that relative fitness is a polynomial of order n. The Appendix provides sufficient conditions for the existence of a stable polymorphic equilibrium and derives a closed form solution for the equilibrium share of type T under those conditions. The model can be easily extended to include institutions and examples of crowding in or our for $n > 1$ can be found without difficulty.

5.3 Degenerate priors and hardwired preferences

As already mentioned, our approach rests on two implicit assumptions. First, evolution does not endow individuals with perfect priors. Second, evolution does not provide in-

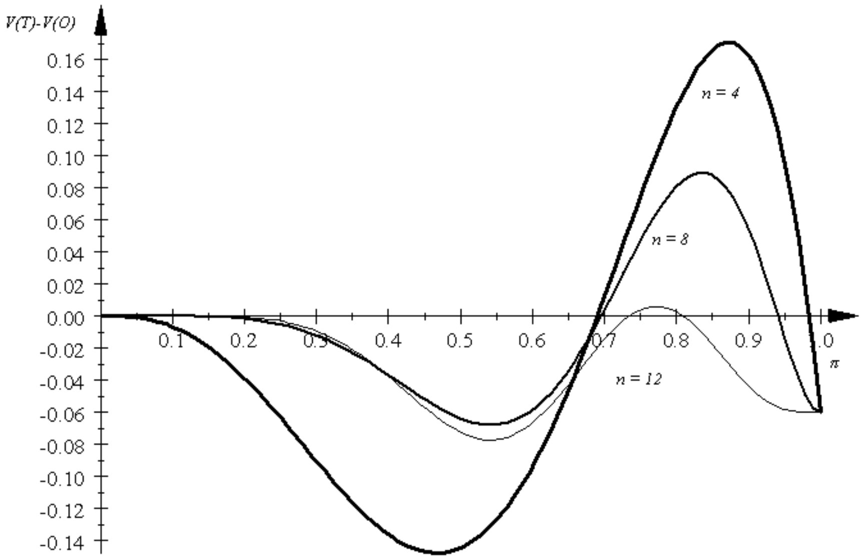

Figure 5: Relative fitness against share of type T for $n = 4$, $n = 8$, $n = 12$. Other parameters are as follows: $\theta = 4.8$, $\alpha = 3.6$, $\rho = 0.06$.

dividuals with direct preferences over trusting behavior. Consider the first assumption. If evolution were completely free to endow individuals with any prior, it could in principle choose a prior distribution assigning probability one to the actual value of π. An individual endowed with this prior and with preferences defined over material outcomes would always trust if and only if it were materially optimal to do so, thus gaining an evolutionary advantage. On the other hand, given that all relevant information is already encapsulated in the prior, this individual would have no scope (and no reason) for learning. Such an individual would thus be immune from the consensus effect. However, it is clear (and well accepted in the evolution of preferences literature) that evolution may be constrained in its ability to endow individuals with priors that are so accurate. In their literature survey, for instance, Robson and Samuelson (2010) argue that "the complexity of a perfect prior is simply out of reach of a trial and error mutation process" (see also Samuelson and Swinkels, 2006 for related discussions). In practice, beliefs usually reflect collected information as well as in-built prior information.

With the second assumption, we focus on the case where the trusting decision is cognitive and determined by expected material outcomes, rather than by direct utility (or disutility) associated to the act of trusting itself. Notice that the first assumption alone would not achieve much without the second. Whether individuals have perfect priors or not is immaterial if evolution can shape preferences in a way that induces them to trust whenever it is (materially) optimal to do so. In this case, beliefs simply do not matter, since nature can redress any deviation from materially optimal behavior due to imperfect information by operating on direct preferences over actions. However, the result that information has no bearing on behavior sounds obviously implausible. Moreover, there are reasons why the decision of whether to trust should not be hardwired. For instance, when dealing with an acquaintance, people usually benefit from assessing the information available about the acquaintance's past record before choosing whether to trust or not. An individual with hardwired preferences for trusting (or not trusting) would be unable to do this. On the other hand, the distinction between strangers and acquaintances is to some extent a matter of degree (the amount of information available) rather than kind. As a result, it is not clear why nature should hardwire behavior in one case but not in the other.

6 Contributions to the existing literature and future work

Our work focuses on the implications of the consensus effect for trust and trustworthiness. Although the consensus effect has attracted increasing attention from the empirical literature (notably, the work by Engelmann and Strobel 2000 and 2012, Ellingsen et al. 2010, and several others), there is little theoretical work that builds on it (the exceptions being Goeree and Großer 2006 and Vanberg 2008). This paper contributes to filling this gap. Our starting point is the observation that the consensus effect provides an indirect channel that links an individual's *preferences* (as reflected in his behavior when acting as a second-mover/agent) to his first mover/principal decision via his *beliefs.* We show that this indirect channel generates a selection effect that may induce trustworthy ethical attitudes to persist in the long-run.

In addition to generating this insight, our contribution to the literature is three-fold. First, we show that a polymorphic population (where different ethical attitudes coexist) may be stable. The presence of heterogeneity in behavior has been extensively documented by the empirical literature. In his survey of experimental economics, Samuelson (2005) argues that "Perhaps one of the most robust findings to emerge from experimental economics is that such heterogeneity is widespread and substantial. Despite this, heterogeneity has often not played a prominent role in many theoretical models." In our framework, heterogeneity emerges endogenously, as an equilibrium feature. As we have seen, the complementarities and substitutabilities generated by the consensus and the selection effects imply that trustworthy ethical attitudes may fall short of spreading all the way within the population. In the polymorphic equilibrium, the population contains a positive fraction of both trustworthy and opportunistic individuals. Our theoretical analysis is one of the few to account for heterogeneity in behavior.

Second, our argument for the persistence of trustworthy behavior does not rely on the observability of other players' preferences, and this stands in contrast with previous literature, where observability of other players' preferences is generally considered necessary for non-materialistic behavior to emerge spontaneously.[25] An exception to this is Huck and Oechssler (1998). They consider a setup where players cannot recognize their opponent's type. However, differently from our model, players observe the composition of the population from which the opponent is drawn. Type-dependent beliefs (a necessary ingredient for the consensus effect) are therefore ruled out. Gamba (2012) allows for type-dependent beliefs, although (in contrast with our setup) she does not build on the consensus effect. Her model relies on a different mechanism from ours and cannot accommodate the persistence of heterogeneity.

Third, our analysis shows that once the consensus effect is factored in, the relationship between ethical attitudes and institutions becomes quite complex. Better institutions may end up *crowding out* trustworthy ethical attitudes. The fact that policies may sometimes affect preferences in ways that counteract the very purpose for which the policies where designed has been noted elsewhere in the literature, notably by Bohnet et al. (2001) and Bar-Gill and Fershtman (2004 and 2005), although in contexts different to ours. Our

[25] When preferences are observable, 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 et al., 2007, and also Samuelson, 2001, for a discussion of conceptual problems related to the observability of preferences). Alternative explanations for the emergence of pro-social preferences include kin-selection/assortative matching arguments – see Alger and Weibull (forthcoming) and Weibull and Salomonsson (2006) for recent contributions. See also van Veelen (2006) for implications of these types of explanations.

paper presents a novel rationale for why such crowding out may occur.

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 many sequential social dilemmas. Consider for instance the implications of the consensus effect in an ultimatum game, where individuals may be reciprocal or opportunistic. When acting as a proposer, a reciprocal individual believes it more likely that the responder is reciprocal, and is thus more inclined to behave generously towards the responder than an opportunistic individual. If the share of reciprocal types in the population is sufficiently high, that is the optimal action to take. This generates a potential advantage for reciprocal individuals that may offset the disadvantage they suffer when acting as responders – since reciprocal agents penalize themselves by rejecting ungenerous offers. The implication is that reciprocal types may spread.[26] A similar argument may apply within the context of the gift exchange game, the control game, the sequential public good game or the sequential prisoner's dilemma. In sum, we believe that our observations apply to a variety of games. Future work should be devoted to clarify this intuition.

7 Appendix

7.1 Proof of Lemma 1

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

$$h(\pi|x,\tau_P = T) = \pi \frac{g(x|\pi)f(\pi)}{\int_{z\in\mathcal{P}} zg(x|z)dF(z)} = \frac{\pi\tilde{g}(\pi|x)}{E(\pi|x)} \tag{33}$$

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

$$h(\pi|x,\tau_P = O) = (1-\pi)\frac{g(x|\pi)f(\pi)}{\int_{z\in\mathcal{P}}(1-z)g(x|z)dF(z)} = \frac{(1-\pi)\tilde{g}(\pi|x)}{1-E(\pi|x)} \tag{34}$$

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 $b(x,\tau_P) \equiv \Pr(\tau_A = T|x,\tau_P)$ the probability assessment that the agent is

[26] Whether they can spread all the way or not is an open question. However, the rationale we have provided for the existence of a polymorphic equilibrium should apply also in that framework.

of type T made by a type τ_p principal. A type T principal believes that the agent is a type T with probability

$$b(x,T) = \frac{\int_{\pi\in\mathcal{P}} \pi^2 d\tilde{G}(\pi|x)}{E(\pi|x)} = E(\pi|x) + \frac{Var(\pi|x)}{E(\pi|x)} \tag{35}$$

The same probability for a type O principal is

$$b(x,O) = \frac{\int_{\pi\in\mathcal{P}} \pi(1-\pi) d\tilde{G}(\pi|x)}{1-E(\pi|x)} = E(\pi|x) - \frac{Var(\pi|x)}{1-E(\pi|x)}. \tag{36}$$

□

7.2 Stable populations and institutions

In this section we provide proofs for a slightly more general statement of Propositions 1 and 2 of Section 3. In particular, the results are proved for all $\phi \geq 0$. Propositions 1 and 2 can accordingly be seen as special cases (for $\phi = 0$) of Propositions 3 and 4 proved below. We also provide a full characterization of the crowding in and crowding out results.

As in Section 3.2, let $X^\tau_\phi \subseteq X$ denote the set of realizations of x for which type τ chooses to participate, i.e.

$$X^\tau_\phi = \{x \in X : b(x,\tau) \geq R(\phi)\}. \tag{37}$$

where $R(\phi) \equiv [\alpha - \theta\phi]/[\theta(1-\phi)]$. Relative fitness now becomes

$$\begin{aligned} V_\pi(T) - V_\pi(O) = \\ \left(\mathcal{G}(X^T_\phi,\pi) - \mathcal{G}(X^O_\phi,\pi)\right)\theta(1-\phi)\left(\pi\frac{\theta-\rho}{\theta} - R(\phi)\right) - \mathcal{G}(X^O_\phi,\pi)\rho(1-\phi). \end{aligned} \tag{38}$$

Proposition 1 can be restated as

Proposition 3. *When objective evidence is coarse $\pi = 0$ is asymptotically stable iff $b(1,T) \geq R(\phi)$.*

Proof. Consider a share $\epsilon > 0$ sufficiently small of type T. A fraction $1-\epsilon$ of the population observes $x = 0$ while a fraction ϵ observes $x = 1$. If $R(\phi) \leq b(0,O)$, then $X^T = X^O = \{0,1\}$. Relative fitness (38) then reduces to $-\rho(1-\phi) < 0$. Suppose then that $R(\phi) > b(0,O)$. If $R(\phi) \leq b(1,O) = b(0,T)$ (which implies $R(\phi) \leq b(1,T)$), then $X^O = \{1\}$ and $X^T = \{0,1\}$. Relative fitness (38) is thus,

$$V_\epsilon(T) - V_\epsilon(O) = (1-\epsilon)\theta(1-\phi)\left(\epsilon\frac{\theta-\rho}{\theta} - R(\phi)\right) - \epsilon\rho(1-\phi) \tag{39}$$

which is negative for $\epsilon > 0$ sufficiently small. Suppose now that $b(1,O) = b(0,T) < R(\phi) \leq b(1,T)$, so that $X^O = \varnothing$ and $X^T = \{1\}$. From (38),

$$V_\epsilon(T) - V_\epsilon(O) = \epsilon\theta(1-\phi)\left(\epsilon\frac{\theta-\rho}{\theta} - R(\phi)\right) \tag{40}$$

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

As for Proposition 2, condition (15) ($\theta/\rho > 2$) remains unchanged. Condition (16) generalizes into

$$R(\phi) < \frac{\theta - 2\sqrt{\rho(\theta-\rho)}}{\theta}, \tag{41}$$

and (17) becomes

$$b(0,O) < R(\phi) \leq b(1,O). \tag{42}$$

Then,

Proposition 4. *When objective evidence is coarse, conditions (15), (41), and (42) are necessary and sufficient for the existence of an asymptotically stable polymorphic population. Under these conditions the stable share of type T is*

$$\pi^* \equiv \frac{\theta(1+R(\phi)) - 2\rho + \Delta_\phi}{2(\theta-\rho)} \in (0,1), \tag{43}$$

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

Proof. We start off by showing that $b(0,O) < R(\phi) \leq b(1,O)$ is necessary for a stable polymorphic population. Suppose first that $b(0,O) \geq R(\phi)$. This implies that all individuals participate independently of their type or the signal x they observe. Since $X^T = X^O = \{0,1\}$, relative fitness (38) is strictly negative for all $\pi \in (0,1)$. As a result, no stable polymorphic population exists. Suppose then that $b(1,O) < R(\phi)$. This implies $X^O = \varnothing$ so that type O never participate. Relative fitness (38) is thus given by

$$V_\pi(T) - V_\pi(O) = \theta(1-\phi)\mathcal{G}(X^T,\pi)\left[\pi\frac{\theta-\rho}{\theta} - R(\phi)\right] \tag{44}$$

where $\mathcal{G}(X^T,\pi)$ is equal to one if $X^T = \{0,1\}$, is equal to π if $X^T = \{1\}$, and is equal to zero if $X^T = \varnothing$. Inspection of (44) 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 $b(0,O) < R(\phi) \leq b(1,O)$, $X^O = \{1\}$ and $X^T = \{0,1\}$ (since $b(1,T) \geq b(0,T) = b(1,O)$). From (38), relative fitness is thus given by,

$$V_\pi(T) - V_\pi(O) = (1-\pi)\theta(1-\phi)\left(\pi\frac{\theta-\rho}{\theta} - R(\phi)\right) - \pi\rho(1-\phi) \tag{45}$$

Clearly enough, any stable polymorphic population, if there is any, is a root of the RHS of (45) (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 (45) has real roots in* $(0,1)$ *if and only if* $\theta > 2\rho$ *and* $R(\phi) < \frac{\theta-2\sqrt{\rho(\theta-\rho)}}{\theta}$. *The roots are* π^* *(as given by 43) and* $\hat{\pi} = \pi^* - \Delta_\phi/(\theta-\rho)$. *Both* π^* *and* $\hat{\pi}$ *are in the interval* $(R(\phi),1)$.

Inspection of (45) reveals that the RHS is an increasing-decreasing function taking negative values for $\pi = 0$ and $\pi = 1$. If $R(\phi) \geq (\theta-\rho)/\theta$, then $V_\pi(T) - V_\pi(O) < 0$ for all $\pi \in (0,1)$ so that the RHS of (45) has no real root in $(0,1)$. However, if $R(\phi) < (\theta-\rho)/\theta$ and $\sqrt{\theta^2(1-R(\phi))^2 - 4\rho(\theta-\rho)} > 0$, then the RHS of (45) has two real roots,

$$\hat{\pi}, \pi^* = \frac{\theta(1+R(\phi)) - 2\rho \pm \sqrt{\theta^2(1-R(\phi))^2 - 4\rho(\theta-\rho)}}{2(\theta-\rho)}. \tag{46}$$

of which the largest is π^* as given by (43). It is immediate to check that $\hat{\pi} < \pi^* < 1$, and that $\pi^* > \hat{\pi} > 0$ requires $R(\phi) > (2\rho-\theta)/\theta$. To sum up, $R(\phi)$ must satisfy

$$\frac{2\rho-\theta}{\theta} < R(\phi) < \min\left\{\frac{\theta-2\sqrt{\rho(\theta-\rho)}}{\theta}, \frac{\theta-\rho}{\theta}\right\}, \tag{47}$$

Notice that

$$\frac{2\rho-\theta}{\theta} < \frac{\theta-2\sqrt{\rho(\theta-\rho)}}{\theta} \tag{48}$$

only if $\theta > 2\rho$. Moreover, given $R(\phi) \geq 0$ for all $\phi \in [0, \alpha/\theta]$ and $\theta > 2\rho$, the first inequality in (47) always holds and can be ignored. Under $\theta > 2\rho$ we also have

$$\frac{\theta-\rho}{\theta} > \frac{\theta-2\sqrt{\rho(\theta-\rho)}}{\theta} \tag{49}$$

Hence, $R(\phi) < \frac{\theta-2\sqrt{\rho(\theta-\rho)}}{\theta}$ and $\theta > 2\rho$, are necessary and sufficient conditions for the RHS of (45) to have two real roots in the interval $(0,1)$. Moreover $\theta > 2\rho \Rightarrow \pi^* > \hat{\pi} > R(\phi)$. This proves Lemma 2.

The proof of Proposition 4 is concluded by noticing that $d(V_z(T) - V_z(O))/dz|_{z=\hat{\pi}} \geq 0$ and $d(V_z(T) - V_z(O))/dz|_{z=\pi^*} < 0$, so that $\hat{\pi}$ is unstable and π^* is stable. □

The crowding in and crowding out results immediately follow from Propositions 3 and 4.

Corollary 1 (*crowding in*)

a) Assume $\theta > 2\rho$ and consider two institutional environments ϕ' and ϕ'' with $\phi'' > \phi'$. Then, a positive fraction π^* of trustworthy individuals is possible under ϕ'' but not under ϕ' if

$$b(0,O) < R(\phi'') < \min\left\{b(1,O), \frac{\theta - 2\sqrt{\rho(\theta-\rho)}}{\theta}\right\} < R(\phi'), \tag{50}$$

b) Consider two institutional environments ϕ' and ϕ'' with $\phi'' > \phi'$. If conditions (15), (41), and (42) are satisfied for both ϕ' and ϕ'', then the stable share of trustworthy π^* is larger under ϕ'' than under ϕ'.

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

$$R(\phi'') < b(0,O) < R(\phi') < \min\left\{b(1,O), \frac{\theta - 2\sqrt{\rho(\theta-\rho)}}{\theta}\right\}, \tag{51}$$

7.3 Sufficient conditions for a stable polymorphic equilibrium with a generic number n of signal realizations

Suppose that $\alpha/\theta \in (n/(n+3), (n+1)/(n+3))$. This implies $X^O = \{n\}$ and $X^T = \{n-1, n\}$. Relative fitness (31) then becomes,

$$V_\pi(T) - V_\pi(O) = \pi^{n-1}\left\{n(1-\pi)(\pi\theta - \alpha) - \rho\pi(n+1-n\pi)\right\} \tag{52}$$

If $\Delta(n) \equiv \sqrt{[n(\theta-\alpha)) - \rho(1+n)]^2 - 4n\rho\alpha}$ is a real number, then the expression in graphs has two real roots given by

$$\hat{\pi} = \frac{n(\theta+\alpha) - \rho(1+n) - \Delta(n)}{2n(\theta-\rho)}, \; \pi^* = \frac{n(\theta+\alpha) - \rho(1+n) + \Delta(n)}{2n(\theta-\rho)} \tag{53}$$

Assume then $[n(\theta-\alpha) - \rho(1+n)]^2 > 4n\rho\alpha$. It is immediate to show that $\hat{\pi} < \pi^* < 1$. On the other hand, $\pi^* > \hat{\pi} > 0$ requires $n\theta > (1+n)\rho$. [This follows since $n\theta > (1+n)\rho \Rightarrow n(\theta+\alpha) - (n+1)\rho > \Delta(n) \Rightarrow \hat{\pi} > 0 \Leftrightarrow \pi^* > 0$. Notice that $\hat{\pi} > 0$ is not only sufficient for $\pi^* > 0$, but also necessary. This follows from inspection of the quadratic expression in graphs in (52) which is increasing/decreasing and negative for $\pi = 0$ and $\pi = 1$.] To sum up, sufficient conditions for a polymorphic equilibrium for a generic n are

1. $\alpha/\theta \in (n/(n+3), (n+1)/(n+3))$.

2. $[n(\theta-\alpha) - \rho(1+n)]^2 > 4n\rho\alpha$.

3. $n\theta > (n+1)\rho$.

Notice that for all $n \geq 1$ there always exists $\rho_n > 0$ such that conditions 2 and 3 are satisfied for all $\rho \in (0, \rho_n)$. For each n, fix a value for α such that

$$\alpha \in \left(\frac{n}{n+3}\theta, \frac{n+1}{n+3}\theta \right) \tag{54}$$

so that condition 1 is also satisfied for all $\theta \in (\alpha, \infty)$. It is then clear that for any n there exists an open set of parameter values such that a stable polymorphic equilibrium exists.

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] Bowles, S. (1998) "Endogenous preferences: The cultural consequences of markets and other economic institutions" Journal of Economic Literature, 36: 75-111.

[12] Butler, J., Giuliano, P., and Guiso, L. (2009) "The right amount of trust" Mimeo, UCLA.

[13] 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.

[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] Dufwenberg, M., and Kirchsteiger, G. (2004) "A theory of sequential reciprocity" Games and Economic Behavior 47: 268-298.

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

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

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

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

[21] Fehr, E. and S. Gächter (2002) "Do incentive contracts undermine voluntary cooperation?" Institute for Empirical Research in Economics, Zurich University, Working Paper No. 34.

[22] 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

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

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

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

[26] Geanakoplos, J., Pearce, D., and Stacchetti, E., (1989) "Psychological games and sequential rationality" Games and Economic Behavior, 1: 60-79.

[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] Rabin, M. (1993) "Incorporating Fairness into Game Theory and Economics" American Economic Review. 85: 1281-1302.

[35] Robson, A.J. (1990) "Efficiency in evolutionary games: Darwin, Nash, and the secret handshake" Journal of Theoretical Biology 144: 379-396.

[36] 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.

[37] 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.

[38] Samuelson, L. (2001) "Introduction to the evolution of preferences" Journal of Economic Theory. 97: 225-230.

[39] Samuelson, L. (2004) "Information-based relative consumption effects" Econometrica. 72: 93-118.

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

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

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

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

[44] Tabellini, G., (2008) "The scope of cooperation: values and incentives", Quarterly Journal of Economics, 123: 905-950.

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

[46] van Veelen, M. (2006) “Why kin and group selection models may not be enough to explain human other-regarding behaviour”, Journal of Theoretical Biology 242: 790-797.

[47] Weibull, J.W., and M. Salomonsson (2006) “Natural selection and social preferences”, Journal of Theoretical Biology 239: 79-92.