Procedural Rationality and Equilibrium Trust

Robert Rowthorn
Rajiv Sethi

SFI WORKING PAPER: 2007-01-002

Procedural Rationality and Equilibrium Trust*

Robert Rowthorn† Rajiv Sethi‡

January 13, 2007

Abstract

This paper examines determinants of the steady state distribution of trusting behavior in a population of principals and agents, where the former learn from experience in accordance with boundedly rational procedures. For any given distribution of agent types, the long run distribution of principal behavior is characterized. It is shown that heterogeneity in the behavior of principals persists under both the sampling procedure (Osborne and Rubinstein, 1998) and the maximum average procedure (Rustichini, 2003). For the sampling procedure, we identify sufficient conditions under which greater resistance to control on the part of agents results in greater equilibrium trust among principals. We also show that the maximum average procedure, despite its greater sophistication, can result in poorer performance than the sampling procedure both from the perspective of the principal, and also with respect to aggregate payoffs.

*This paper was written while both authors were visiting the Santa Fe Institute. We thank the Institute, and especially Sam Bowles, for their hospitality, and Carlos Rodriguez for comments on an earlier version. We also thank two referees and the Editor for suggesting a number of improvements.

†Faculty of Economics, University of Cambridge (rer3@econ.cam.ac.uk).

‡Department of Economics, Barnard College, Columbia University (rs328@columbia.edu).

1 Introduction

Many situations require us to choose the extent to which our fate is left in the hands of others. We can give others substantial discretion over actions that affect our well-being, or we can constrain their choices in ways that leave us less vulnerable. A willingness to leave others unconstrained signals *trust* in the sense that we expect them to act in a manner reasonably congruent with our interests, even if this entails some material sacrifice on their part. For the same reason, constraining the discretion of others signals distrust in their willingness to put adequate weight on our well-being. Such decisions can be difficult because of the possibility that signals of distrust may themselves induce self-interested behavior in others, while signals of trust may elicit generous responses. Trust is riskier but can also be more rewarding than distrust.

The British welfare state provides a good example of the dilemmas which arise in the context of trust. Doctors, academics and school teachers in the public sector were for a long time regarded as responsible professionals who could be trusted to perform their jobs conscientiously with little outside monitoring. This has changed considerably in recent years. A large bureaucracy has arisen which monitors the performance of such professionals and specifies, often in great detail, what they should do and how they should allocate their time. This development is widely resented because conformity with the rules is time-consuming and restricts the freedom of those concerned to follow their professional judgment. It is also resented as manifestation of distrust. Many professionals believe that extensive monitoring undermines professional commitment and encourages a mercenary attitude (Le Grand, 2006). The defenders of the present monitoring system, including many economists, dismiss such complaints as self-serving or exaggerated. They point to various scandals that have occurred and to the failure of many professionals to perform adequately when left to their own devices. They acknowledge that the present monitoring system does provoke a negative behavioral response in some professionals, but believe this is more than offset by its benefits in identifying rogue elements and raising the minimum standard.

The idea that trust is often rewarded by generosity on the part of those who are trusted is also familiar in the experimental economics literature (Berg, Dickhaut, and McCabe, 1995). One consequence of this is that contracts based on explicit economic incentives (which signal distrust) can interfere with norms of trust and reciprocity and hence result in lower payoffs to principals than certain implicit contracts (see, for instance, Frey 1993, 1997, Gneezy and Rustichini 2000a, 2000b, and Fehr and Gächter, 2002). A recent paper by Falk and Kosfeld (2006) reveals in a particularly clear manner both the risks and the potential rewards involved in the exercise of trust. Their experiment involves a principal and an agent, the latter of whom is given an endowment. Part of this endowment can be transferred to the principal, who receives a multiple (greater than one) of the transferred amount. The principal has the option of restricting the set of possible transfers available to the agent by setting a lower bound below which the transfer cannot lie. Clearly a principal who expects agents to behave selfishly will choose to restrict the transfer. On the other hand, the choice to restrict transfers can be interpreted as a signal of distrust, to which certain agents may respond by lowering their transfers. Falk and Kosfeld report that a majority of agents do, in fact, respond in this way: choosing *higher* transfers when left unrestricted than when restricted (such agents are referred to as *control-averse*). Moreover, a majority of principals appear to anticipate such behavior, and thus choose to leave transfers unrestricted. The average returns to trust are significantly higher than those to distrust, even though a sizeable minority of agents transfer nothing when trusted.

How might the decision of whether or not to trust be made in practice? It is reasonable to suppose that it is based on some combination of experimentation, experience, and habit. When individuals respond positively to trust and negatively to distrust, they are increasing the expected return to trust, and encouraging the spread of trusting behavior in the population at large, even when that is not their intention. We model this process and attempt to identify some of the determinants of the incidence of equilibrium trust. Experimental evidence suggests that there is considerable heterogeneity across individuals with respect to their beliefs regarding whether or not it pays to trust others, and hence heterogeneity also in the extent of trusting behavior. Responses to trust also vary widely, with some individuals behaving selfishly, while others reward trust and still others seem to punish it. We account for the persistence of heterogeneous beliefs and actions using a notion of procedural rationality introduced by Osborne and Rubinstein (1998), and further developed by Rustichini (2003). We show that such procedures result in a non-degenerate equilibrium distribution in the population of the extent of trusting behavior on the part of principals, and examine how changes in the behavior of agents alters this equilibrium distribution. Of particular interest are the kinds of changes in agent behavior that result in greater trust on the part of principals. Intuition suggests that an increase in control-aversion should result in greater equilibrium trust. We show that this is true if control-aversion is sufficiently extreme, but not in general: greater control-aversion can sometimes result in a lower incidence of equilibrium trust.

The sampling procedure is extremely simple: principals try each available action once and adopt whichever one yields the greatest payoff for use in all subsequent periods. The maximum average procedure is somewhat more sophisticated. Each action is initially sampled once, and the one yielding the highest payoff is selected to begin with. After this, the principal selects whichever action has resulted in the highest average payoff to date. This procedure can involve multiple switches back and forth between actions over time. Despite its greater sophistication, however, it turns out that the maximum average procedure can result in lower expected payoffs for the principal, as well as lower payoffs in the aggregate, relative to the sampling procedure. The conditions under which this occurs are economically interesting, and are satisfied in the Falk and Kosfeld data.

In this paper we take the distribution of agent behavior to be given and focus on the equilibrium behavior of principals under boundedly rational procedures. This leaves open the question of why agents might behave in the manner that they do. In recent work, Ellingsen and Johannesson (2006) construct an elaborate model of signaling which may be used to rationalize the behavior of both principals and agents. The logic of their argument (applied to the Falk and Kosfeld experiment) is roughly as follows. Individuals are heterogeneous with respect to both preferences and beliefs, and care not only about monetary payoffs but also about the extent to which they are regarded as generous by others. Generous principals are more likely to believe that they have been matched with generous agents, and generous agents prefer to make larger transfers when faced with generous principals. In a fully separating equilibrium, selfish principals control and generous ones do not. Selfish agents transfer only what they are forced to, and generous agents transfer less when controlled (believing correctly that the principal is selfish) than when trusted. Principals in this model are fully rational and, conditional on their own type, have identical beliefs about the agents with whom they are matched. In our framework, on the other hand, principals who are *ex-ante* identical can end up taking different actions in equilibrium as a result of different sampling histories.

The term "trust" has been used in many different ways in academic discourse. We use the word to describe an action that will only be beneficial to the principal if the agent is sufficiently unselfish. We also assume that the decision to trust is influenced by the beliefs of the principal regarding the nature of the agent, and that these beliefs are themselves based on prior experience. Hence our notion of trust has both "calculative" and "personal" components in the sense of Williamson (1993). Trust is not simply a label for a decision that leaves an individual vulnerable to actions of another party. It is a belief that this vulnerability will not be exploited for personal gain.

2 Evidence

Consider the following simple interaction between a principal and an agent (Falk and Kosfeld, 2006). The agent has an endowment of e units and chooses a transfer of w units to the principal. The resulting payoffs are $e - w$ to the agent and βw to the principal, where $\beta > 1$. Prior to the transfer, the principal may set a lower bound c below which w cannot lie. A principal who expects agents to maximize monetary returns will choose to restrict the transfer (thereby ensuring a payoff of βc instead of 0).

Falk and Kosfeld group agents into three classes based on their responses. Selfish agents transfer the minimum feasible amount in either case, which is c when transfers are restricted and 0 otherwise. Inequity-averse agents make transfers that are strictly positive and insensitive (subject to feasibility) to whether or not they are trusted by the principal. Finally, control-averse agents make transfers that are strictly *higher* when left unrestricted relative to the case in which they are restricted. Although there is a fair amount of heterogeneity within groups, there is enough clustering of individual responses to make the classification meaningful. When $e = 120$, $\beta = 2$, and $c = 10$, the median transfers of the three agent types are as follows.

	Control	Trust
Selfish	10	0
Inequity-Averse	12	12
Control-Averse	10	30

(1)

Modal transfers reveal a similar pattern: selfish agents transfer the minimum possible, inequity-averse agents make transfers that are independent of control, and control-averse agents transfer much more when trusted than when controlled. A few agents cannot be classified in either one of the three categories: they transfer more when controlled (as in the case of selfish agents) but do not always choose the minimum possible transfer. We shall refer to such agents as *manipulation-averse.* Their specific motivation in the Falk and Kosfeld experiment is unknown, but it is concievable that they were reacting negatively to trust because they believed that the principal had chosen this option, not out of respect for them, but in order to maximize his own payoff. In fact, it is entirely possible that a portion of the agents who are characterized as selfish by Falk and Kosfeld are in fact manipulation-averse.

Interestingly, the beliefs of principals who choose to control are systematically less optimistic (regarding anticipated agent transfers) than those of principals who trust. Principals tend to choose the action that they expect will lead to the greatest transfer, but there is considerable heterogeneity (and hence inaccuracy) in beliefs. A central purpose of this paper is to provide an account of the manner in which such belief heterogeneity and inaccuracy may arise and persist over time as agents learn from experience and choose actions in accordance with well-specified but boundedly rational procedures.

3 The Sampling Procedure

Consider a large population of principals and agents, where the latter can be partitioned into the three groups identified above. For simplicity, suppose that there is no behavioral heterogeneity *within* groups of agents, and that agent transfers are as follows:

	Control	Trust
Selfish	c	0
Inequity-Averse	x_e	y_e
Control-Averse	c	y_t

Assume that

$$0 < c < x_e = y_e < y_t. \tag{2}$$

and that the population shares of the three agent types are given by $s = (s_1, s_2, s_3)$. Principals are aware of the actions available to them, and the possible payoff consequences of these actions, but are unaware of the distribution of agent types. It is assumed that they form beliefs based on experimentation and experience. Osborne and Rubinstein (1998) introduce the following simple procedure for doing this: each principal samples each of the two actions exactly once, and chooses the one yielding a higher payoff thereafter. Ties are broken with uniform probability. The three possible rankings of actions then arise with the following probabilities:

$$\begin{aligned}\Pr(\pi_c > \pi_t) &= s_1\\ \Pr(\pi_c = \pi_t) &= s_2^2\\ \Pr(\pi_c < \pi_t) &= 1 - s_1 - s_2^2\end{aligned}$$

Let the behavior of principals be represented by $\sigma = (\sigma_1, \sigma_2)$, where σ_1 is the proportion of principals choosing to control, while σ_2 is the proportion choosing to trust. Then $\sigma^* = (\sigma_1^*, \sigma_2^*)$ is said to be a *sampling equilibrium* if the likelihood with which action i is chosen under the sampling procedure is precisely equal to σ_i^*. Since s is exogenously given, there is a unique sampling equilibrium given by:

$$(\sigma_1^*, \sigma_2^*) = \left(s_1 + \frac{1}{2}s_2^2, 1 - s_1 - \frac{1}{2}s_2^2\right). \tag{3}$$

Any sampling equilibrium can be viewed as a steady state of a dynamic process involving a large population with entry and exit (Sethi, 2000). In general games, with players in all positions sampling simultaneously, there can exist multiple steady states, some of which may be unstable. In the simple case considered here, however, the sampling equilibrium (3) is both unique and stable. As long as all three types of agents are present in the agent population, there will be persistent heterogeneity in the behavior of principals. Note that there is an asymmetry in our treatment of principals and agents, and it is only the former who are actively engaged in learning. This may be justified in the present context since agents, when called upon to move, are fully aware of the choice made by the principal and hence the payoff consequences of any action that they might take.

This simple model can be used to address the effects on trust and efficiency of changes in the distribution of agent types. Three types of changes are possible: (i) shifts from selfishness to inequity-aversion, (ii) shifts from selfishness to control-aversion, and (iii) shifts from inequity-aversion to control-aversion. Consider each of these in turn. A shift from selfishness to inequity-aversion corresponds to an increase in s_2 at the expense of s_1. From (3), this raises the long-run incidence of trust. The same effect occurs when there is a shift from selfishness to control-aversion (an increase in s_3 at the expense of s_1), or from inequity-aversion to control-aversion (an increase in s_3 at the expense of s_2). These effects are intuitive: greater selfishness among agents tends to erode trust over time while a high degree of control-aversion tends to encourage trust among principals. Whether or not this holds under more general specifications of agent behavior is explored in the section to follow.

Since principals are unaware of the distribution of agent types, they do not know whether trust or control has the higher expected payoff. In fact, trust will be the more rewarding strategy on average if

$$s_3 \left(y_t - c \right) > s_1 c, \tag{4}$$

and the less rewarding strategy if the inequality is reversed. The expected transfer from agents to principals in equilibrium is

$$T(s) = \sigma_1^* \left(\left(1 - s_2 \right) c + s_2 x_e \right) + \sigma_2^* \left(s_2 y_e + s_3 y_t \right).$$

Since $\beta > 1$, any transfer is of greater benefit to the principal than its cost to the agent. Thus, summing across principals and agents, the expected aggregate payoff is monotonically increasing in T. Do increases in the equilibrium incidence of trust correspond to higher transfers on average (and hence a greater aggregate payoff)? The following example shows that this need not be the case.

Example 1. Suppose $s = (0.5, 0.2, 0.3)$ and $s' = (0.5, 0.1, 0.4)$. Then $\sigma^*(s) = (0.520, 0.480)$ and $\sigma^*(s') = (0.505, 0.495)$. If $c = 10$, $x_e = y_e = 15$, and $y_t = 40$, then $T(s) = 12.92 < 13.96 = T(s')$. However, if $x_e = y_e = 35$ instead, then $T(s) = 16.92 > 15.97 = T(s')$.

In this example, a shift from inequity-aversion to control-aversion (s to s') causes the average transfer to rise for one combination of x_e and y_e, while exactly the same shift causes the average transfer to fall for higher values of x_e and y_e. To understand why, note that the shift from inequity-aversion to control-aversion has a direct and an indirect effect. The direct effect causes expected payoffs to decline when principals control and rise when they trust. The indirect effect occurs because principals alter their behavior in response to the shift in payoffs, thereby raising the proportion of those who trust. This indirect effect helps to raise the average transfer. Even so, as the case $x_e = y_e = 35$ indicates, the average transfer can decline with the shift in agent preferences if inequity-averse agents are sufficiently generous. Under these conditions, the shift in agent preferences means that a large additional penalty is imposed on agents who choose control. The resulting loss outweighs the additional payoffs due to the fact that trust is more widespread and more profitable than before, causing the average transfer to fall.

A shift away from selfishness towards either inequity-aversion or control-aversion, however, results in an unambiguous increase in transfers provided condition (4) is satisfied. The reason is that expected transfers under control cannot fall, while those under trust rise. This is accompanied by an increase in the proportion of principals who trust. Given (4), overall transfers must therefore rise. The robustness of this finding to more general distributions of agent preferences is explored next.

4 Generalized Agent Preferences

We now consider the implications of the sampling procedure for an arbitrary distribution of agent types. Selfish and inequity-averse types share in common the feature that they transfer at least as much when controlled as they do when trusted. We also allow for the existence of manipulation-averse agents. As mentioned above, such agents transfer more when they are controlled than when they are trusted, and the amount they transfer when controlled is more than the stipulated minimum.

We shall refer to selfish, inequity-averse, and manipulation-averse types collectively as *control-tolerant.* Control-averse agents, in contrast, transfer strictly greater amounts under trust than under control. The transfers by agents in the two cases (control and trust) are shown below.

	Control	Trust
Selfish $x_i = c > 0,\ y_i = 0$	x_1	y_1
Inequity-Averse $y_i > 0, x_i = \max(c, y_i)$	x_2	y_2
	$\vdots$	$\vdots$
	x_j	y_j
Manipulation-Averse $x_i > \max(c, y_i)$	x_{j+1}	y_{j+1}
	$\vdots$	$\vdots$
	x_m	y_m
Control-Averse $x_i \geq c > 0,\ y_i > x_i$	x_{m+1}	y_{m+1}
	$\vdots$	$\vdots$
	x_{m+n}	y_{m+n}

These transfers have the following structure: for $i = 1, ..., m$, we have $x_i \geq y_i$, while for $i = m+1, ..., m+n$, we have $x_i < y_i$. Note that we allow the possibility for there to exist control-averse types who transfer *less* than some control tolerant types when trusted. We also allow for the possibility that there exist control-averse types who transfer *more* than some control tolerant types when controlled. Neither of these possibilities can arise in the simple model considered in the previous section.

As before, let $s = (s_1, ..., s_{m+n})$ denote the preference distribution in the agent population, and let $\sigma = (\sigma_1, \sigma_2)$ denote the probabilities with which the sampling procedure leads the principal to select control and trust respectively. Let $I = \{1, ..., m+n\}$ and define the sets $L(i)$ and $E(i)$ as follows:

$$\begin{aligned} L(i) &\equiv \{j \in I \mid x_i > y_j\}, \\ E(i) &\equiv \{j \in I \mid x_i = y_j\}. \end{aligned}$$

Then the likelihood of selecting control is as follows:

$$\sigma_1(s) = \sum_{i \in I} s_i \left(\sum_{j \in L(i)} s_j + \frac{1}{2} \sum_{j \in E(i)} s_j \right). \tag{5}$$

Now consider a shift in agent preferences from s to s', such that the share of some control-tolerant type falls and the share of some control-averse type rises by the same amount. Formally, suppose that there exists at that point $k \in \{1, ..., m\}$, $l \in \{m+1, ..., m+n\}$ and $\delta > 0$ such that $s'_k = s_k - \delta$, $s'_l = s_l + \delta$, and $s'_i = s_i$ for all $i \notin \{k, l\}$. We shall refer to any such change in the agent preference distribution as a shift from control-tolerance to control-aversion.

By definition, we must have $x_k \geq y_k$ and $x_l < y_l$, but it is entirely possible that either $y_l < y_k$ or $x_l > x_k$. In other words, the agents whose preferences change from control-tolerant to control-averse may in fact become *less* generous under trust or *more* generous under control. This considerably complicates the analysis of the manner in which such shifts alter the equilibrium incidence of trust under the sampling procedure. As the following example shows, even though a shift from control-tolerance to control-aversion always raises the expected returns to trust, it can result in a *lower* incidence of steady state trust under the sampling procedure.

Example 2. Suppose that the agent population consists of the following types:

	Control	Trust
Inequity-Averse	20	20
Control-Averse	10	15
Control-Averse	18	25

By definition

$$\sigma_1(s) = \Pr(\pi_c > \pi_t) + \frac{1}{2}\Pr(\pi_c = \pi_t) = (1 - s_2)\, s_2 + \frac{1}{2}s_1^2$$

Let $s = (0.2, 0.1, 0.7)$ and $s' = (0.1, 0.2, 0.7)$. Then $\sigma_1(s) = 0.110 < 0.165 = \sigma_1(s')$.

Example 2 shows that an increase in the proportion of control-averse types in the population can raise the equilibrium incidence of control under sampling. The following result establishes this cannot occur if all control-averse agents make the minimum required transfer when distrusted. This condition means that control-averse agents always punish a controlling principal to the maximum possible extent.

Proposition 1. *Suppose $y_i \geq c$ for all $i \in \{j+1, ..., m\}$ and $x_i = c$ for all $i \in \{m+1, ..., m+n\}$. Then any shift in the agent preference distribution from control-tolerance to control-aversion results in a lower equilibrium incidence of control (and hence a higher incidence of trust).*

Proof: See Appendix.

Notice that when there are no manipulation-averse types the proposition only requires the simpler condition that $x_i = c$ for all $i \in \{m+1, ..., m+n\}$. In this case, if all control-averse types punish distrust to the maximal permissible degree, then an increase in the proportion of such types will increase the equilibrium incidence of trust. In each of the treatments considered by Falk and Kosfeld, the median agent response when controlled was to choose the minimum feasible transfer. This suggests that control-averse types do, in fact, frequently punish distrust to the maximal possible degree.

While the analysis above assumes that each available action is sampled just once before one of them is adopted, one could easily extend this to the case in which k actions are sampled, corresponding to Osborne and Rubinstein's notion of $S(k)$ equilibrium. As k approaches infinity it is easily seen that the resulting equilibrium involves optimal choice on the part of principals: the payoff maximizing action will be chosen with certainty. For any finite k, however, both actions will be selected with positive probability.

One shortcoming of the sampling procedure (even for $k > 1$) is that the choices of individuals are not sensitive to accumulating payoff experience once the initial selection has been made. An alternative is the *maximum average procedure* which selects at each stage the action with the highest average payoff based on all accumulated experience to date (Rustichini, 2003). We next show that the heterogeneity of behavior on the part of principals arises also in this case.

5 The Maximum Average Procedure

A principal adopting the maximum average procedure begins by sampling each action once, and subsequently chooses whichever action has resulted in the highest average payoff to date. Ties are broken with uniform probability. This can result in periodic switching of actions as a string of poor outcomes lowers the average return to the incumbent action. Consider, for instance, the example of three agent types with payoffs as given in (1) and an agent population composition $s = (0.3, 0.5, 0.2)$. In this example, depending on the experience of the principal, switching backwards and forwards between trust and control sometimes occurs. One particular realization of the process is shown in Figure 1. However, as we demonstrate below, every principal will eventually settle on one particular course of action, either control or trust. We also prove that there is a non-zero probability that the eventual choice will be control and a non-zero probability that it will be trust. In this example, the expected returns to control and trust are 11 and 12 respectively. Since there is a non-zero probability that the principal will settle down choosing control, which has the lower expected return, the maximum average procedure is not therefore efficient.

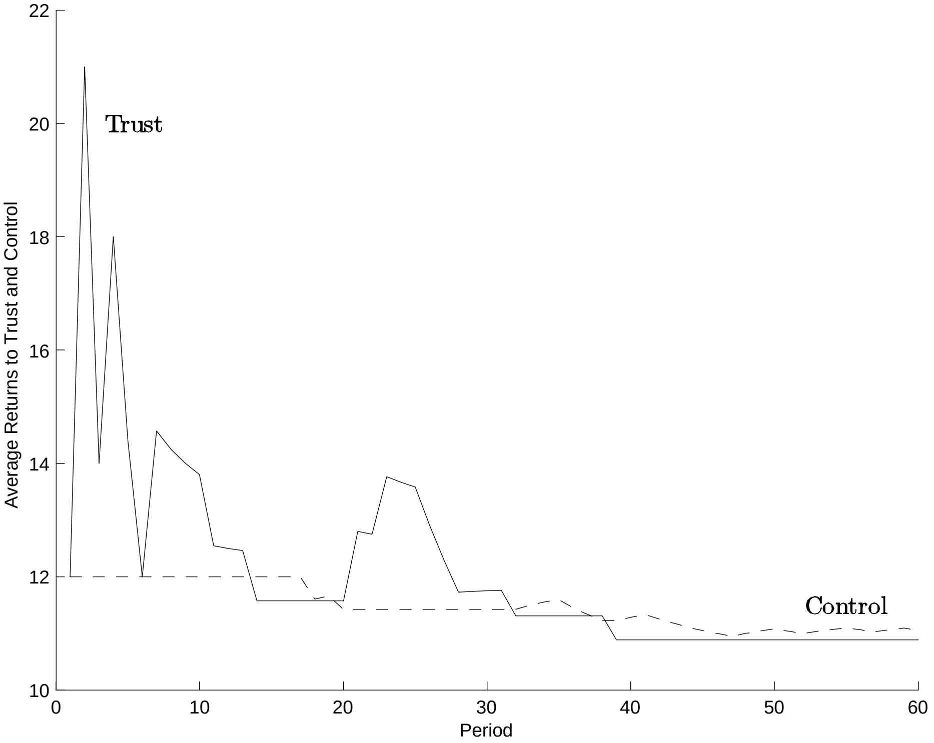

Figure 1. Switching under the Maximum Average Procedure

Let ρ_i^* denote the probability that the maximum average procedure selects action i (we show below that these probabilities are well-defined and that $\rho_1^* + \rho_2^* = 1$). While closed form solutions are difficult to obtain, we can use numerical methods to explore the manner in which the probabilities depend on the distribution of agent behavior. The following example applies the maximum average procedure to the agent distributions used in Example 1.

Example 3. Suppose $s = (0.5, 0.2, 0.3)$, $s' = (0.5, 0.1, 0.4)$, $c = 10$, and $y_t = 40$. If $x_e = y_e = 15$, then $\rho^*(s) = (0.71, 0.29)$, $\rho^*(s') = (0.62, 0.38)$, and $T(s) = 12.15 < 13.16 = T(s')$. However, if $x_e = y_e = 35$ instead, then $\rho^*(s) = (0.59, 0.41)$, $\rho^*(s') = (0.56, 0.44)$, and $T(s) = 16.62 > 15.58 = T(s')$.

Example 1 illustrated how, under the sampling procedure, a shift from inequity-aversion to control-aversion (s to s') may cause the average transfer to rise for one combination of x_e and y_e, while exactly the same shift causes the average transfer to fall for higher values of x_e and y_e. Example 3 illustrates the same point using the maximum average procedure. Comparing these examples reveals an important difference between the two procedures. In each case, the equilibrium proportion of principles who trust, and hence the average transfer, is lower under the maximum average procedure than under the sampling procedure. Since a principal who chooses trust enjoys a higher expected return than one who chooses control, this result indicates that the maximum average procedure may be less efficient than the sampling procedure. This result was established using simulation. In the following section we use more formal methods to compare the two procedures.

Example 2 showed that, even though a shift from control-tolerance to control-aversion always raises the expected returns to trust, it may result in a *lower* incidence of steady state trust under the sampling procedure. The following example illustrates the same point for the maximum average procedure.

Example 4. Suppose that agent types are exactly as in Example 2, and let $s = (0.2, 0.1, 0.7)$ and $s' = (0.1, 0.2, 0.7)$. Then $\rho_1(s) = 0.08 < 0.11 = \rho_1(s')$.

In this particular case, the maximum average procedure and the previously discussed sampling procedure lead to same outcome: an increase in the population share of control-averse types results in a higher equilibrium incidence of control (and hence a lower incidence of trust).

It remains to be shown that the maximum average procedure does indeed select one of the actions, and that each of the actions can be selected with positive probability. Let $x(u)$ denote the return that the principal receives on the uth occasion that he chooses to control. Since the agent population is fixed, these returns are identically distributed and serially independent random variables with mean $\bar{x}$ and variance σ_x^2. The maximum and minimum values that the variables can take are denoted by $x_{\max}$ and $x_{\min}$ respectively. Similarly, $y(v)$ is the return that the principal receives on the vth occasion that he chooses to trust. Such returns are identically distributed random variables and are serially independent. The minimum value that these variables can take is denoted by $y_{\min}$. They have mean $\bar{y}$ and variance σ_y^2. The following relationships are assumed to hold

$$\begin{aligned}
y_{\min} &< x_{\min} < \bar{y} \\
\bar{x} &< y_{\max} \\
\bar{x} &\neq \bar{y} \\
\Pr\left(x(u) = x_{\min}\right) &= p > 0 \\
\Pr\left(y(v) = y_{\min}\right) &= q > 0 \\
\Pr\left(y(v) = y_{\max}\right) &= r > 0
\end{aligned}$$

Note that this specification includes that of the previous section as a special case. It is more general since it allows for control-averse types for whom $x(i) > x_{\min}$ when controlled and for inequality-averse types for whom $x(i) > y(i) > 0$.

The average returns from control and trust are simply

$$\begin{aligned}
\bar{x}(u) &= \frac{1}{u}\sum_{i=1}^{u} x(i) \\
\bar{y}(v) &= \frac{1}{v}\sum_{i=1}^{v} y(i)
\end{aligned}$$

The principal begins by choosing trust and control once each, and from then onwards chooses the next step according to which choice has yielded the highest average return to date. If the average returns are equal he chooses between actions with equal probability. Formally, if $w = u+v$, then on round $w+1$ the principal will choose trust if $\bar{x}(u) > \bar{y}(v)$ and control if $\bar{x}(u) < \bar{y}(v)$. If $\bar{x}(u) = \bar{y}(v)$ he will choose either action with probability 0.5. Since $w = u + v$, it is obvious that $u \to \infty$ or $v \to \infty$ as $w \to \infty$. Hence, $\Pr(u \to \infty)+ \Pr(v \to \infty) = 1$. In addition, we have:

Proposition 2. $\Pr(u \to \infty) > 0$, $\Pr(v \to \infty) > 0$, *and* $\Pr(u \to \infty$ *and* $v \to \infty) = 0$.

Proof: See Appendix.

The above proposition implies that with probability 1 the principal will eventually stop switching between trust and control and settle for one or the other. The probability that he will settle for control is $\rho_1^* = \Pr(u \to \infty)$, and the probability that that he will settle for trust is $\rho_2^* = \Pr(v \to \infty)$. Both of these are nonzero, which means that the maximum average procedure does not converge to an efficient solution. The result is persistent heterogeneity in the behavior of principals for a wide range of agent behaviors. In this respect the maximum average procedure behaves qualitatively in much the same manner as the sampling procedure.

6 Performance of Procedures

The maximum average procedure is clearly more sophisticated that the sampling procedure, and appears also to be more sensible. Nevertheless, as shown above, there exist instances in which a principal adopting the former will obtain a smaller expected payoff than a principal adopting the latter. In this section we show that for the benchmark) model discussed in section 3, the maximum average procedure selects control with strictly higher probability that the sampling procedure does, provided that inequity-averse agents are not too common and selfish agents are not too rare. This happens even when trust is the action which yields the higher expected payoff.

Suppose that, as before, there are three agent types with population shares (s_1, s_2, s_3), who respond to control and trust as follows:

	Control	Trust
Selfish	c	0
Inequity-Averse	x_e	y_e
Control-Averse	c	y_t

and that

$$0 < c < x_e = y_e < y_t$$

as before. Then the maximum average procedure selects control with higher probability than the sampling procedure selects control:

Proposition 3. For any given $\eta > 0$, there exists $\varepsilon > 0$ such that, if $s_1 \geq \eta$ and $s_2 < \varepsilon$ then $\rho_1^* > \sigma_1^*$.

Proof: Let $\eta > 0$ be given and assume that $s_1 \geq \eta$. From (3) we have

$$\sigma_1^* = s_1 + \frac{1}{2}s_2^2. \tag{6}$$

The maximum average procedure begins by sampling each action once. If the payoff 0 is realized when trust is sampled, then control is selected by the procedure regardless of all future payoff realizations. This event occurs with probability s_1. Now consider the following event. When the maximum average procedure is applied, the payoffs obtained at the first stage are c when control is sampled and y_e when trust is sampled (so trust is initially selected). This is followed by n realizations of the payoff 0, where n is the largest integer strictly below y_e/c. Clearly such an integer exists and is at least equal to 1. Such a sequence is possible since the first $n-1$ such realizations maintain the average payoff to trust at or above c. The last realization pushes the average payoff to trust strictly below c, after which the principal switches to control and stays there for all future periods. The probability of this event is $s_2(1-s_2)s_1^n/2$ if y_e/c is an integer, and $s_2(1-s_2)s_1^n$ otherwise. Since the two events are mutually exclusive, we have

$$\rho_1^* \geq s_1 + \frac{1}{2}s_2(1-s_2)s_1^n \geq s_1 + \frac{1}{2}s_2(1-s_2)\eta^n. \tag{7}$$

Define ε as follows

$$\varepsilon = \frac{\eta^n}{\eta^n + 1}.$$

Clearly $\varepsilon > 0$ and satisfies $\varepsilon = (1-\varepsilon)\eta^n$. For any $0 < s_2 < \varepsilon$, we therefore have

$$s_2 < \epsilon = (1-\varepsilon)\eta^n < (1-s_2)\eta^n,$$

and hence from (6) and (7), $\sigma_1^* < \rho_1^*$. ■

Note that this result holds regardless of whether or not trust is the more rewarding action on average. Suppose that the agent distribution is such that condition (4) is satisfied: $s_3\left(y_t - c\right) > s_1 c$. Then trust yields a higher expected payoff than control. Despite this, the maximum average procedure selects control with higher probability than the sampling procedure, provided that s_2 is sufficiently small relative to s_1. Note that since transfers from agent to principal are multiplied by $\beta > 1$, the maximum average procedure in this case results not just in poorer outcomes from the perspective of the principal, but also gives rise to smaller aggregate payoffs. Despite being considerably simpler and seemingly less reasonable, the sampling procedure results here in superior performance along both of these dimensions.

7 Conclusions

There is considerable heterogeneity across principals in their willingness to trust, as well as heterogeneity in the manner in which agents respond to trust. Since the returns to trust are fully determined by the distribution of agent behavior, this implies heterogeneity in the payoffs of principals, and suboptimality in the choices of at least some subset of principals. Such heterogeneity persists here because principals have limited information and are boundedly rational. Learning from experience leads *ex-ante* identical individuals to make different choices *ex-post.* Somewhat surprisingly, the naive sampling procedure can generate outcomes that are superior both for the principal and in the aggregate relative to the more demanding maximum average procedure.

Intuition suggests that an increase in control-aversion on the part of agents should result in less control (and hence greater trust) on the part of principals in equilibrium. This need not be the case (under either procedure) unless agent behavior is sufficiently punitive in the face of distrust. One consequence of this finding is that efforts to increase the incidence of trust in society require not only greater aversion to distrust in the response of agents, but also a rather extreme and highly punitive form of control-aversion.

Appendix

Proof of Proposition 1. Consider a shift in agent preferences from s to s', such that $s'_k = s_k - \delta$, $s'_l = s_l + \delta$, and $s'_i = s_i$ for all $i \notin \{k, l\}$, where $k \in \{1, ..., m\}$, $l \in \{m+1, ..., m+n\}$, and $\delta > 0$. Since $s'_i = s_i$ for all $i \notin \{k, l\}$, any difference between $\sigma_1(s)$ and $\sigma_1(s')$ arises from the possibility of sampling an individual who has switched from type k to type l. The likelihood of drawing such an individual on any given trial is δ. Let such individuals (formerly type k, currently type l) be referred to as being of type z. Consider the following three possibilities: (i) a type z individual is drawn both under control and under trust, (ii) a type z individual is drawn under control but not under trust, and (iii) a type z individual is drawn under trust but not control. Let Δ_1, Δ_2, and Δ_3 be the difference between $\sigma_1(s)$ and $\sigma_1(s')$ that can be attributed to these three events respectively. We need to show that

$$\sigma_1(s) - \sigma_1(s') \equiv \Delta_1 + \Delta_2 + \Delta_3 > 0.$$

First suppose that a type z individual is drawn on both trials, which occurs with probability δ^2. Since $x_k \geq y_k$ this event would have led to the choice of control with probability at least $\frac{1}{2}$ at s. Since $x_l < y_l$, it leads to control with probability 0 at s'. Hence

$$\Delta_1 \geq \frac{1}{2}\delta^2 > 0. \tag{8}$$

Next consider the event in which a type z individual is drawn under control but not under trust. This occurs with probability $\delta(1-\delta)$. Since $x_l = c \leq x_k$, the likelihood that trust is chosen at s' is at least as high as the likelihood that trust is chosen at s. Hence $\Delta_2 \geq 0$.

Finally consider the event in which a type z individual is drawn under trust but not under control. This occurs with probability $\delta(1-\delta)$. If $y_l \geq y_k$, then the likelihood that trust is chosen at s' is at least as high as the likelihood that trust is chosen at s. Hence $y_l \geq y_k$ implies $\Delta_3 \geq 0$. Since $\Delta_2 \geq 0$ and $\Delta_1 > 0$, we therefore have $\sigma_1(s) > \sigma_1(s')$ as required.

If $y_k < y_l$, there may exist types i such that $x_i \in [y_l, y_k]$. In this case,

$$\Delta_3 = -\delta(1-\delta)\left(\sum_{x_i \in (y_l, y_k)} s_i + \frac{1}{2}\sum_{x_i \in \{y_l, y_k\}} s_i\right) \tag{9}$$

Since type k is control-tolerant and type l is control-averse $x_k \geq y_k$ and $y_l > x_l$. Hence if $y_k < y_l$

$$x_l \leq x_k. \tag{10}$$

In this case

$$\Delta_2 = \delta(1-\delta)\left(\sum_{y_i \in (x_l, x_k)} s_i + \frac{1}{2}\sum_{y_i \in \{x_l, x_k\}} s_i\right). \tag{11}$$

The proposition assumes that for control-averse types the transfer when they are controlled is exactly equal to c. This implies that $y_l > x_l = c$. Thus, for any type i with $x_i \in [y_l, y_k]$,the strict inequality $x_i > c$ must hold. Such a type cannot be control-averse or selfish and must therefore be inequity-averse or manipulation-averse. Either way, $y_i \leq x_i$.Since $y_k \leq x_k$ and $x_i \leq y_k$ it follows that

$$y_i \leq x_k \tag{12}$$

The proposition assumes that for manipulation-averse types the transfer under trust is at least equal to c. This is also the case for inequity-averse types. Hence,

$$x_l = c \leq y_i \tag{13}$$

From (12) and (13) it follows that $y_i \in [x_l, x_k]$. Using this fact together with (11) and (9), we get:

$$\begin{aligned}
\Delta_2 &= \delta\left(1-\delta\right)\left(\sum_{y_i \in \left(x_l, x_k\right)} s_i + \frac{1}{2}\sum_{y_i \in \left\{x_l, x_k\right\}} s_i\right) \\
&\geq \delta\left(1-\delta\right)\left(\sum_{x_i \in \left(y_l, y_k\right)} s_i + \frac{1}{2}\sum_{x_i \in \left\{y_l, y_k\right\}} s_i\right) = -\Delta_3,
\end{aligned}$$

so $\Delta_2 + \Delta_3 \geq 0$. Since $\Delta_1 > 0$ from (8), we have $\sigma_1\left(s\right) - \sigma_1\left(s'\right) > 0$ as required. ■

Proof of Proposition 2. To establish that $\Pr(u \to \infty) > 0$, consider any sequence that begins with an arbitrary $x(1)$, followed by $y(1) = y_{\min}$. Such a sequence occurs with probability q. Since $\bar{y}(1) = y_{\min} < x_{\min} \leq x(1) = \bar{x}(1)$, the principal will switch back to control on the third round. On every subsequent round the principal will also choose control since $\bar{y}(1) = y_{\min} < x_{\min} \leq \bar{x}(u)$ for all u. The result of the initial choice will be an infinite sequence $x(1), y_{\min}, x(2), ...$in which $\bar{y}(1) < \bar{x}(u)$ at every stage. Thus, $Pr(u \to \infty) \geq q > 0$.

To establish that $\Pr(v \to \infty) > 0$ consider any infinite sequences $\{x(u)\}, \{y(v)\}$ whose elements are independent and have the assumed distributions. We claim that for any ε, $\eta > 0$ there exists T such that

$$\begin{aligned}
\Pr(|\bar{x}(u) - \bar{x}| &\geq \varepsilon \text{ for all } u > T) \leq \eta \qquad (14) \\
\Pr(|\bar{y}(v) - \bar{y}| &\geq \varepsilon \text{ for all } v > T) \leq \eta \qquad (15)
\end{aligned}$$

This can be proved as follows. The variances of $\bar{x}(u)$ and $\bar{y}(v)$ are $\sigma^2_{\bar{x}(u)} = \sigma^2_x/u$ and $\sigma^2_{\bar{y}(v)} = \sigma^2_y/v$ respectively. Using Tchebychev's inequality we can show that

$$\begin{aligned}
\Pr\left(|\bar{x}(u) - \bar{x}| > \varepsilon/2\right) &\leq \left(\frac{2}{\varepsilon}\right)^2\left(\frac{\sigma_x^2}{u}\right) \\
\Pr\left(|\bar{y}(v) - \bar{y}| > \varepsilon/2\right) &\leq \left(\frac{2}{\varepsilon}\right)^2\left(\frac{\sigma_y^2}{v}\right)
\end{aligned}$$

Choosing T such that $\eta\varepsilon^2 T/4 > \max\left(\sigma_x^2, \sigma_y^2\right)$ yields (14–15). These inequalities imply

$$\begin{aligned}
\Pr(\bar{x}(u) &> \bar{x} - \varepsilon \text{ for all } u > T) > 1 - \eta \\
\Pr(\bar{x}(u) &< \bar{x} + \varepsilon \text{ for all } u > T) > 1 - \eta \\
\Pr(\bar{y}(v) &> \bar{y} - \varepsilon \text{ for all } v > T) > 1 - \eta \\
\Pr(\bar{y}(v) &< \bar{y} + \varepsilon \text{ for all } v > T) > 1 - \eta
\end{aligned}$$

Let the events D, E, F be defined as follows

$$\begin{aligned}
D &: \bar{y}(v) > \bar{y} - \varepsilon \text{ for all } v > T \\
E &: \bar{y}(v) = y_{\max} \text{ for } v \leq T \\
F &: \bar{y}(v) > \bar{y} - \varepsilon \text{ for all } v
\end{aligned}$$

It is clear that

$$\Pr(D \mid E) \geq \Pr(D).$$

From above, $\Pr(D) > 1 - \eta$. For all k it is also the case that $\Pr\left(y(k) = y_{\max}\right) = r$, and hence $\Pr(E) = r^T$. Since $y_{\max} > \bar{y}$ it follows that $F \supset D \cap E$ and hence using Bayes' Rule

$$\begin{aligned}\Pr(F) &\geq \Pr(D \cap E) \\ &= \Pr(D \mid E)\Pr(E) \\ &= \Pr(D)\Pr(E) \\ &> (1-\eta)r^T\end{aligned}$$

Thus,

$$\Pr\left(\bar{y}(v) > \bar{y} - \varepsilon \text{ for all } v\right) > (1-\eta)r^T > 0$$

Choose $\varepsilon = \bar{y} - x_{\min} > 0$. With the appropriate T it follows that

$$\Pr\left(\bar{y}(v) > x_{\min} \text{ for all } v\right) > (1-\eta)r^T > 0$$

Hence

$$\begin{aligned}&\Pr\left(\bar{x}(1) = x_{\min} \text{ and } \bar{y}(v) > x_{\min} \text{ for all } v\right) \\ =\ &\Pr\left(\bar{x}(1) = x_{\min}\right)\Pr\left(\bar{y}(v) > x_{\min} \text{ for all } v\right) \\ >\ &p(1-\eta)r^T > 0\end{aligned}$$

where $\Pr\left(\bar{x}(1) = x_{\min}\right) = \Pr\left(x(1) = x_{\min}\right) = p > 0$. This establishes that there is a positive probability that after the first pair of trials, the principal will stick with trust forever. Hence $\Pr(v \to \infty) > 0$.

The only thing that remains is to establish that a path cannot spend an infinite time in *both* trust and control with positive probability. Let $\varepsilon = \left|\bar{y} - \bar{x}\right|/3 > 0$. Suppose that $\bar{y} > \bar{x}$ and consider a hypothetical path which contains an infinite subsequence of trust. Along this path, for suitable T,

$$\begin{aligned}&\Pr\left(\bar{y}(v) > \bar{x}(u) \text{ for all } u, v > T\right) \\ \geq\ &\Pr\left(\bar{y}(v) > \bar{y} - \varepsilon \text{ for all } v > T\right)\Pr\left(\bar{x}(u) < \bar{x} + \varepsilon \text{ for all } u > T\right) \\ >\ &(1-\eta)^2\end{aligned}$$

Thus, the probability of choosing control at any time after T is less than $1 - (1-\eta)^2$. By making η sufficiently small and choosing an appropriate value of T , we can make this expression as small as we like. Thus, if a path contains an infinite subsequence of trust, the probability that it also contains an infinite subsequence of control is equal to zero. This establishes that $\Pr(u \to \infty\ \&\ v \to \infty) = 0$ in the case that $\bar{y} > \bar{x}$.

For the case $\bar{x} > \bar{y}$ we proceed as follows. Consider a hypothetical path which contains an infinite subsequence of control. Along this path, for suitable T,

$$\begin{aligned}\Pr(\bar{y}(v) &< \bar{x}(u) \text{ for all } u, v > T) \\ &\geq \Pr\left(\bar{y}(v) < \bar{y} - \varepsilon \text{ for all } v > T\right)\Pr\left(\bar{x}(u) > \bar{x} + \varepsilon \text{ for all } u > T\right) \\ &> (1-\eta)^2\end{aligned}$$

Thus, the probability of choosing trust at any time after T is less than $1 - (1-\eta)^2$. By making η sufficiently small and choosing an appropriate value of T, we can make this expression as small as we like. Thus, if a path contains an infinite subsequence of control, the probability that it also contains an infinite subsequence of trust is equal to zero. This establishes that $\Pr(u \to \infty\ \&\ v \to \infty) = 0$ in the case that $\bar{x} > \bar{y}$. ■

References

[1] Berg, J., J. Dickhaut, and K. McCabe (1995). "Trust, Reciprocity, and Social History" *Games and Economic Behavior* 10: 122-142

[2] Ellingsen and Johannesson (2006). "Pride and Prejudice: The Human Side of Incentive Theory" CEPR Discussion Paper 5678.

[3] Falk, A. and M. Kosfeld (2006). "The Hidden Cost of Control." American Economic Review

[4] Fehr, E. and S. Gächter (2002). "Do Incentive Contracts Crowd Out Voluntary Cooperation?" Institute for Empirical Research in Economics, University of Zürich, Working Paper No. 34.

[5] Frey, B.S. (1993) "Does monitoring increase work effort? The rivalry with trust and loyalty," *Economic Inquiry* 31, 663-670.

[6] Frey, B.S. (1997) *Not just for the money*, Edward Elgar, Cheltenham.

[7] Gneezy, U. and Rustichini, A. (2000a). "A fine is a price," *Journal of Legal Studies* 29, 1-17.

[8] Gneezy, U. and Rustichini, A. (2000b). "Pay enough or don't pay at all," *Quarterly Journal of Economics* 115 (2), 791-810.

[9] Le Grand, J. (2006). *Motivation, Agency, and Public Policy: Of Knights and Knaves, Pawns and Queens,* Oxford University Press, Oxford.

[10] Osborne, M.J. and A. Rubinstein (1998). "Games with Procedurally Rational Players" *American Economic Review* 88: 834-847.

[11] Rustichini, A. (2003). "Equilibria in Large Games with Continuous Procedures" *Journal of Economic Theory* 111: 151-171.

[12] Sethi, R. "Stability of Equilibria in Games with Procedurally Rational Players." *Games and Economic Behavior* 32: 85-104.

[13] Williamson, O. (1993). "Calculativeness, Trust and Economic Organization," *Journal of Law and Economics* 36: 453-486.