DISCUSSION PAPER SERIES

IZA DP No. 2293

Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes

Dirk Sliwka

September 2006

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes

Dirk Sliwka
University of Cologne
and IZA Bonn

Discussion Paper No. 2293
September 2006

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 2293
September 2006

ABSTRACT

Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes*

An explanation for motivation crowding-out phenomena is developed in a social preferences framework. Besides selfish and fair or altruistic types a third type of agents is introduced: These 'conformists' have social preferences if they believe that sufficiently many of the others do too. When there is asymmetric information about the distribution of preferences (the `social norm'), the incentive scheme offered or autonomy granted can reveal a principal's beliefs about that norm. High-powered incentives may crowd out motivation as pessimism about the norm is conveyed. But by choosing fixed wages or granting autonomy the principal may signal trust in a favorable social norm.

JEL Classification: M52, J33, D23

Keywords: social preferences, incentives, intrinsic motivation, motivation crowding-out, social norms, trust, conformity, selection

Corresponding author:

Dirk Sliwka
University of Cologne
Herbert-Lewin-Str. 2
50931 Köln
Germany
E-mail: dirk.sliwka@uni-koeln.de

* I thank Christian Grund, Christine Harbring, Bernd Irlenbusch, Matthias Kräkel, Tom McKenzie, Georg Nöldeke, and Kathrin Pokorny as well as seminar participants in Amsterdam, Bonn, Cologne, Mannheim, Tübingen, Rotterdam, and Zürich for helpful comments and suggestions.

1 Introduction

Economists tend to believe that incentive contracts are beneficial when most aspects of performance are measurable as they make employees work harder. Indeed there are some recent empirical studies on single firms showing that incentive contracts have raised productivity significantly.[1] However, descriptive evidence on the limited overall frequency of use of pay-for-performance schemes may call for more caution.[2] Indeed there seem to be very different views in individual firms on whether contracts based on individual performance are beneficial or not. Whereas some see incentive contracts as an important component of their human resource management practices others take a much more sceptical view and even consider extrinsic incentives harmful.[3]

Psychologists have for quite some time also taken a more sceptical view of extrinsic incentives. Since the work by Deci (1971), it has often been pointed out that monetary incentives can be harmful as they may crowd-out *intrinsic motivation*. Numerous experimental studies have been conducted by psychologists on this issue producing somewhat mixed evidence.[4]

But recently also economic experiments have raised doubts on this issue. In laboratory experiments, for instance, Fehr and Gächter (2002), Irlenbusch and Sliwka (2003), Fehr and Rockenbach (2003) and Falk and Kosfeld (forthcoming) have observed that the ability to set incentives or a restriction of an agent's choice set made principals worse off in contrast to theoretical predictions. Gneezy and Rustichini (2000a) found that weak monetary incentives led to reduced performance outcomes as compared to pure fixed compensation for tasks such as collecting for a charity.

But how can these results be reconciled with the economics of incentives? Kreps (1997) offers an informal discussion of the topic and points out that understanding these issues involves activities unfamiliar to economists but concludes that "messy or not, they are important and must be pursued".

[1] See for instance Lazear (2000) and the overviews provided by Gibbons (1997) or Prendergast (1999).

[2] Parent (2002) for instance surveys different samples of the US working population and finds that at most one quarter of all employees receive some form of compensation based on individual performance. See also Parent and MacLeod (1999).

[3] See for instance Baron and Kreps (1999) Chapters 3 and 11.

[4] Frey and Jegen (2001) and Kunz and Pfaff (2002) review the results of the psychological experiments and the psychologists' theoretical explanations from an economic perspective.

We provide an economic explanation for motivation crowding-out effects based on an extended social preference framework. In recent years a steadily growing economic literature has evolved modeling social preferences, i.e. the way in which individuals care for the well-being of others. Alternative utility functions have been proposed[5] that depart from standard homo oeconomicus assumptions. Many applications of these models quite successfully explain experimentally observed phenomena by assuming that two different types of agents exist in the population: some are *strictly selfish* while others are *fair*, i.e. care to some extent for the well-being of others.

We extend this by introducing a third group of agents who are influenced in their 'moral convictions' by what they think others will do. We assume that such *conformists* will be fair if and only if they think that a sufficiently high fraction of the other *steadfast* agents is fair as well. In this way we model the importance of social norms for individual decisions.

We investigate a basic framework in which a principal can choose whether to *control* or to *trust* an agent and afterwards the agent can exert effort on a task. When the principal controls – for instance by setting incentives – she can ensure that even selfish agents exert effort. When she trusts she makes herself more vulnerable as her payoff depends to a larger extent on the agent's type.

But there is uncertainty about the type of the agent and the distribution of types in the population. The agent of course knows his own type but we assume that the principal has superior information about the type distribution due to her experience with previous employees. From this the explanation for a crowding-out effect arises: By choosing to trust the agent the principal can signal her conviction that most people are fair. If this signal is credible, trust may indeed generate trustworthiness on the part of a conformist agent. On the other hand, when controlling the agent, she reveals her pessimism about the social norm and this may lead conformists to become selfish.

Two special cases of the basic framework are analyzed in more detail. In the first, an employer chooses between a fixed wage and an incentive scheme and we show that paying a fixed wage can indeed be a credible signal of trust even when performance-contingent wages would be optimal with symmetric

[5] See for instance Rabin (1993), Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2002).

information about the agent's type. In the second application, we consider the simple setup proposed by Falk and Kosfeld (forthcoming) where the principal can impose a restriction on the agent's effort. It is shown that our framework yields a straightforward theoretical explanation for their striking experimental results.

But the decision to control or trust employees may also have an impact on the attractiveness of the job and therefore may alter the distribution of types in the organization. Therefore we also investigate these employee self-selection effects of trust.

The paper proceeds as follows. Section 2 discusses related literature. Section 3 presents the basic framework. In Section 4 we analyze conditions under which trust is indeed a credible signal and the results from the general framework are then used to investigate two applications. In Section 5 we extend the model to study selection effects of trust. Section 6 contains further extensions. Section 7 concludes.

2 Related Literature

Motivation crowding out has previously also been analyzed economically in Frey (1997) or Frey and Oberholzer-Gee (1997) who allow for the possibility that an agent's disutility of effort is increasing in the monetary reward for this effort. Bénabou and Tirole (2003) assume that agents do not know their costs of effort perfectly but that the principal has additional information about these effort costs. Motivation crowding-out then occurs as the agent believes that the task is tedious when an incentive scheme is offered.[6] Related to this, but from a different strand of the literature are models by Spier (1992) and Allen and Gale (1992). Both show that incomplete fixed-payment contracts may be chosen in equilibrium as the contract offer can reveal information about the underlying technology.[7]

Holmström and Milgrom (1991) show the optimality of fixed wages when an agent has to work on multiple tasks and the outcome of some important

[6] Note that a precondition for their explanation is that an agent might like to perform the task. Our approach can explain motivation crowding-out even when agents dislike exerting effort as for instance in typical principal-agent models and laboratory experiments where effort was represented by a higher monetary expenditure by the agent.

[7] In Spier (1992) a risk-averse principal has superior information about the profitabilty and risk of a technology. In Allen and Gale (1992) a supplier knows more about his ability to distort a verifiable signal of production costs.

task is unverifiable. Bernheim and Whinston (1998) argue that if certain elements of performance cannot be verified, it may be optimal not to specify other elements in the contract which would be verifiable in principle, as this 'ambiguity' allows good performance to be rewarded.

Conformism and adherence to social norms have been studied in different ways in economics. An early paper incorporating social norms in economics is Akerlof (1980). Bernheim (1994) models conformism by assuming that people care for social status determined by others' beliefs about one's own type, which may lead to distorted individual decisions. Lindbeck et al. (1999) assume that the embarrassment of living on public transfer is decreasing in the share of people living from the transfer. In Kandel and Lazear (1992), Huck et al. (2003) or Fischer and Huddart (2005) members of a team suffer a utility loss when their own effort level falls short of that of their co-workers. Empirical evidence on the importance of conformism in organizations has for instance been found by Ichino and Maggi (2000) who observe a significant positive relationship between a job mover's absenteeism and the average absenteeism of his co-workers in a large Italian bank.[8]

3 The Model

A risk-neutral principal employs an agent. The agent's effort generates a payoff π_A for the agent and π_P for the principal. The agent can be one of three different types. First, there are *selfish agents* who care only about their own well-being, hence, the utility function of a selfish type is given by $u_S = U_S(\pi_A)$. Second, there are trustworthy *fair agents* who have some form of a social preference, i.e. they also care to some extent for the principal's payoff such that their utility function is $u_F = U_F(\pi_A, \pi_P)$. We call these two types the *steadfast* agents as their moral convictions are fixed from the outset. The fraction of fair agents among the steadfasts is given by ϕ, but there is uncertainty about this fraction, hence ϕ is drawn from some prior distribution.

But we assume that there is also a third group, which we call the *conformists*. A conformist is someone who is uncertain about the 'appropriate' behavior in a certain situation and therefore is influenced by social norms.

[8] See also Moffitt (1983), Clark (2003) or Stutzer and Lalive (2004) for evidence on the importance of social norms for the behavior of the unemployed.

If for instance a conformist harms someone else to gain a personal advantage he will suffer from remorse only if he believes that many others would also feel bad about the harmful action. We model this in the following way: a conformist will have some form of a social preference if and only if he believes that sufficiently many of the other steadfast agents also do. We assume that the utility of a conformist $U_C(\pi_A, \pi_P)$ is equal to $U_F(\pi_A, \pi_P)$ if he believes that the median steadfast agent is fair (i.e. if his conditional expectation on ϕ is larger than $\frac{1}{2}$) and equal to $U_S(\pi_A)$ otherwise.[9]

As an employer the principal will typically have learned more from the behavior of previous or other current employees.[10] For simplicity, we assume that she learns the fraction of fair agents which is either ϕ_L or ϕ_H. We focus on the interesting cases where this signal is informative and would affect the preferences of a conformist given that he is able to infer it in the game, i.e. $\phi_H \geq \frac{1}{2} \geq \phi_L$. The fraction of conformists in the population may not be perfectly known but has mean η according to the common prior expectation.

The timing of the game is as follows: first the principal learns her private signal and decides whether to trust or to control the agent $\tau \in \{T, C\}$. Afterwards, the agent chooses an effort level e which affects the principal's as well as his own material payoff such that $\pi_P = \pi_P(\tau, e)$ and $\pi_A = \pi_A(\tau, e)$. This game is a signaling game as the principal's choice may reveal her private information.

Note that a conformist's action choice will always correspond exactly to either that of a steadfastly selfish or to that of a steadfastly fair type depending on his beliefs about ϕ. Hence, the principal's continuation payoff which we denote by Π depends only on the principal's decision τ and on whether the agent acts fairly (then $\Pi = \Pi_{F\tau}$) or selfishly ($\Pi = \Pi_{S\tau}$).

We will first derive a general result and then consider two special cases within this framework. However, these applications share common properties which we use to derive the key result at the outset:

Property 1: $\Pi_{FT} > \Pi_{ST}$ and $\Pi_{FC} \geq \Pi_{SC}$

Property 2: $\Pi_{SC} - \Pi_{ST} > \Pi_{FC} - \Pi_{FT}$

[9] Hence, the game is in a very simple way a psychological game in the sense of Geanakoplos et al. (1989), as players' payoffs are not only affected by what they do but also by what they believe.

[10] For instance large firms have software systems that track performance across different locations. In the subsection 6.2 we show that this can easily be endogenized in a two-period version of the model with multiple agents.

Property 3: $\Pi_{SC} > \Pi_{ST}$ and $\Pi_{FC} \geq \Pi_{FT}$

The first property defines the key characteristic of a trustworthy agent: The principal is always better off when the agent acts fairly. The second captures the idea that the essence of control is to protect against an agent's shirking behavior: the returns to controlling are larger when agents are selfish.[11] The third just characterizes the interesting cases: we want to investigate whether trust may be beneficial even when control is a dominant strategy when the agent's types are known.[12]

4 Trust as a Credible Signal of a Social Norm

4.1 The Existence of a Separating Equilibrium

Of course, in some situations the fairness of an agent does not matter too much for the principal, for instance when control is very effective. When $\Pi_{SC} > \Pi_{FT}$ the principal earns more from a controlled selfish agent than from a trusted trustworthy agent. Then clearly trust can never be optimal even when it favorably affects conformists' behavior.

However, when this is not the case, it may be attractive for a principal to choose trust as this may signal her conviction that most agents are fair. But note that she has to trade off two effects against each other: On the one hand, she will be better off when conformists become trustworthy. But on the other hand, there are also steadfastly selfish agents around and these agents will exert lower effort levels when being trusted.

Trust will be a credible signal when an "optimistic" principal who has received a high signal trusts and a "pessimistic" principal prefers to control the agent. We have to check whether the principal has an incentive to follow this strategy when the agents believe that trust is indeed a credible signal. If ϕ is the principal's subjective probability that a steadfast agent is

[11] For instance, Nagin et al. (2002) found in their study on call center agents that cheating behavior of those employees who have positive attitudes towards the employer varies less with the monitoring rate. The study is also an interesting example of a firm which has superior knowldege of the cheating behavior of their agents by making control calls.

[12] The main result given in Proposition 1 and its applications in Propositions 2 and 3 do not rely on the second part of property 3.

trustworthy, her expected profit when controlling ($\tau = C$) is then given by

$$\underbrace{(1-\eta)(1-\phi)}_{\text{steadfastly selfish agents}} \cdot \Pi_{SC} + \underbrace{\eta}_{\text{conformist turned selfish}} \cdot \Pi_{SC} + \underbrace{(1-\eta)\,\phi}_{\text{steadfastly fair agents}} \cdot \Pi_{FC}. \tag{1}$$

When she trusts, she makes losses from the steadfastly selfish agents as they would work harder when being controlled. But she gains as the conformists become fair. Her expected profits from trusting are

$$\underbrace{(1-\eta)(1-\phi)}_{\text{steadfastly selfish agents}} \cdot \Pi_{ST} + \underbrace{\eta}_{\text{conformist turned fair}} \cdot \Pi_{FT} + \underbrace{(1-\eta)\,\phi}_{\text{steadfastly fair agents}} \cdot \Pi_{FT}. \tag{2}$$

Comparing these two expressions and solving for η yields that the principal will trust when the fraction of conformists is larger than a cut-off value

$$\hat{\eta}(\phi) = 1 - \frac{\Pi_{FT} - \Pi_{SC}}{(\Pi_{FT} - \Pi_{ST}) - \phi(\Pi_{FT} - \Pi_{FC} + \Pi_{SC} - \Pi_{ST})}.$$

This cut-off value is decreasing in ϕ as higher values of ϕ imply a lower expected fraction of steadfastly selfish agents who betray the principal's trust which makes trusting less costly. Using these considerations, we can derive:

Proposition 1 *Given that* $\Pi_{FT} \geq \Pi_{SC}$ *a separating equilibrium exists in which the principal trusts after he has received the good signal and controls after the bad if and only if the fraction of conformists* $\eta \in [\hat{\eta}(\phi_H), \hat{\eta}(\phi_L)]$ *where* $0 < \hat{\eta}(\phi_H) < \hat{\eta}(\phi_L) \leq 1$.

Proof: See Appendix.

It is important to note that the equilibrium exists even though controlling the agent is a dominant strategy under full information. As we have shown, trust may nonetheless be beneficial as it can be a credible signal of a social norm and therefore affect conformists' behavior.

Note that a precondition for the existence of the equilibrium is that there are neither too few nor too many conformists:[13] when there are only a few conformists even an optimistic principal prefers to control the agents.

[13] In subsection 6.1 all pure strategy equilibria are derived. As is shown there, the separating equilibrium is the unique pure strategy equilibrium in this range when $E[\phi] < \frac{1}{2}$. When $E[\phi] \geq \frac{1}{2}$ it coexists with a pooling equilibrium in which the principal always controls.

The higher the fraction of conformists, the more attractive it is to signal optimism about the social norm. But when there are too many conformists even a pessimistic principal would want to imitate this signal and then of course it would no longer be credible.

4.2 Incentives and Identification

This application is concerned with the possible impact of a compensation scheme on employees' identification with the objectives of a firm. Consider a situation in which an agent always earns a base wage w. A principal can choose whether to give an unconditional wage increase of $\Delta \geq 0$ (then she 'trusts') or introduce a piece rate $\beta \leq 1$ (then she 'controls').[14] The principal's revenue is equal to the effort e exerted by the agent at cost $c(e) = \frac{c}{2}e^2$. Hence, when the principal trusts $\pi_A(T, e) = w + \Delta - c(e)$ and $\pi_P(T, e) = e - w - \Delta$ whereas when she controls $\pi_A(C, e) = w + \beta e - c(e)$ and $\pi_P(C, e) = e(1 - \beta) - w$.

Adopting the terminology of Akerlof and Kranton (2005) the steadfast agents are either selfish 'outsiders' or trustworthy 'insiders'. Selfish outsiders care only for their own well-being and hence $U_S(\pi_A(\tau, e)) = \pi_A(\tau, e)$. Trustworthy insiders, however, also identify to some extent with the well-being of their employer

$$U_F(\pi_A(\tau, e), \pi_P(\tau, e)) = \pi_A(\tau, e) + \mu \cdot \pi_P(\tau, e).$$

The higher μ, the stronger the identification with the objectives of the organization. Whether conformists act as insiders or as outsiders now depends upon their beliefs about the prevailing social norm, i.e. their beliefs about ϕ updated following the observation of the principal's choice of a compensation scheme.

First, it is instructive to consider an insider's objective function when the piece rate has been chosen. She maximizes $w + \beta e - \frac{c}{2}e^2 + \mu((1 - \beta)e - w)$. From the first order condition we obtain the reaction function

$$e = \frac{(1 - \mu)\beta + \mu}{c}.$$

Hence, insiders respond to incentives, but an insider's optimal effort choice

[14] The key result is generalized for the case of an unrestricted choice of a linear incentive scheme in subsection 6.3.

is less sensitive to the power of the incentive scheme β as compared to an outsider's (with $\mu = 0$). It is straightforward to check that properties 1 and 2 are always satisfied. Property 3, which requires that control is preferred when the agent's type is known, holds if the piece rate β is not too large for a given Δ.[15]

From Proposition 1 we can directly infer that a separating equilibrium may indeed exist in which the principal prefers to raise the fixed wage instead of using the piece rate. But for a given fraction η of conformists, we can also use the result to formulate requirements for a salary increase such that it is a credible signal of the firm's confidence that most employees are trustworthy:

Proposition 2 *An optimistic principal can credibly signal trust by raising the fixed salary instead of paying the piece rate β when the salary increase*

$$\Delta \in \left[\frac{\mu(\eta+(1-\eta)(2-\beta)\beta\phi_L)-\beta(1-\beta)}{c}, \frac{\mu(\eta+(1-\eta)(2-\beta)\beta\phi_H)-\beta(1-\beta)}{c} \right].$$

This set is non-empty if $\mu \geq \beta(1-\beta)$ and if the fraction of conformists is sufficiently large.

Proof: See Appendix.

Raising the fixed wage instead of paying the piece rate is costly as selfish outsiders exert less effort. But it can be beneficial as conformists become insiders when the wage increase is a credible signal of the principal's confidence. Signaling this confidence becomes possible as an optimistic principal suffers less from not setting incentives. However, if there are many conformists the signal may not be credible as a pessimistic principal may want to imitate the signal. This effect is prevented when the "costs of trust" are raised by increasing the fixed salary. Hence, it may indeed be the case that performance pay crowds out motivation but performance-independent payments support a "crowding in".

The result is illustrated in Figure 1. The proportion of conformists η is drawn on the abscissa and the wage increase Δ on the ordinate. If Δ is below the upper boundary for a given η, an optimistic principal chooses a fixed wage when this turns conformists into trustworthy insiders. If it is below the lower boundary a pessimistic principal would do the same. The larger the fraction of conformists the more attractive it is to signal trust.

[15] It always holds irrespectively of Δ when $\beta < \frac{1-2\mu}{1-\mu}$.

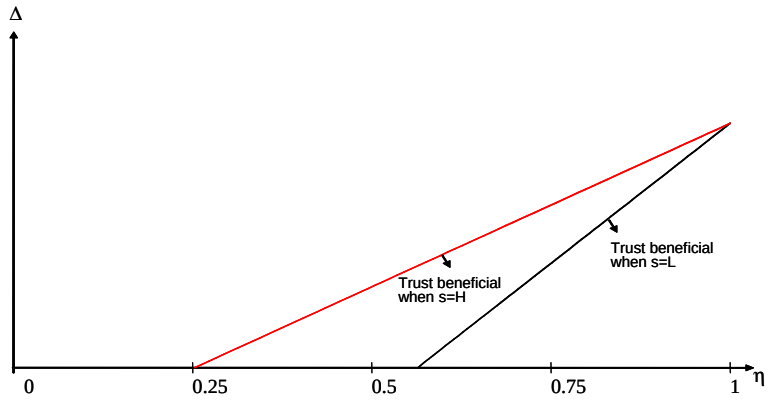

Figure 1: A salary increase as a signal of trust.

Not setting monetary incentives can then only be a credible signal if it is accompanied with a sufficiently large salary increase.[16]

From a more general perspective, the introduction of an incentive scheme has two effects which the employer has to trade-off against each other: it has an *incentive effect* as even employees who do not identify with the objectives of the organization work harder. But there may also be a *crowding-out effect* as it may reveal that not identifying with the goals of the organization is a widespread behavior.

4.3 Trust and Restrictions

The preconditions for our set-up – costly effort choices and uncertainty about the behavior of others – typically hold for most economic laboratory experiments. We therefore apply our framework to a recent experiment by Falk and Kosfeld (forthcoming). In this experiment a principal had the binary choice of whether to impose a lower boundary $r \in \{0, R\}$ on the set of feasible effort levels. An agent then chose an effort $e \in [r, K]$ resulting in a payoff of $2e$ for the principal and $K - e$ for the agent.

If all agents were selfish they would never choose positive effort levels and principals would always impose the restriction. As Falk and Kosfeld already lay out in their paper, principals will also impose a restriction if not all are selfish but some agents care for fairness in the sense that they are

[16] Note that the precondition for the existence of such an equilibrium $\mu \geq \beta(1-\beta)$ holds irrespectively of β when $\mu \geq \frac{1}{4}$.

inequity averse. By imposing a restriction, the principal can protect herself against selfish agents without altering the inequity averse agents' effort choices. But Falk and Kosfeld surprisingly found that average efforts chosen in the experiment with anonymous one-shot interaction were significantly higher when principals did not impose a restriction.

We now also assume that agents can be selfish $U_S(\pi_A) = \pi_A$ or inequity averse with Fehr and Schmidt (1999)-type utility functions such that

$$U_F(\pi_A, \pi_P) = \pi_A - v(\pi_A - \pi_P).$$

In addition we allow for the possibility that agents are conformists who care for fairness if and only if sufficiently many steadfast agents do so. We will show that this yields a straightforward theoretical explanation for the striking experimental result.

A selfish agent will always choose the lowest possible effort level. Without restriction an inequity averse agent chooses[17]

$$e^* = \arg\max_e K - e - v((K - e) - (2e)).$$

As long as $e^* \geq R$ he will do the same when a restriction is imposed. It is straightforward that the game always satisfies the properties 1-3 laid out above. Using Proposition 1 we obtain the following:

Proposition 3 *When $e^* > R$ and if the fraction of conformists is neither too small nor too large a separating equilibrium exists in which the principal imposes no restriction if and only if she received the good signal. In this equilibrium the observed average effort levels are higher when no restriction is imposed.*

Proof: See Appendix.

Hence, not imposing the restriction can be a credible signal of the principal's confidence that many agents are indeed fair: it is costly as unfair agents work less but these costs are smaller when the principal is optimistic

[17] For instance, when an agent becomes "infinitely inequity averse" $e^* \to \frac{K}{3}$ as this equalizes the principal's and agent's payoff. In the baseline experiment many people indeed chose $e = \frac{K}{3} = 40$, which is well above $R = 10$.

and this may make signaling possible. Hence, our approach yields a possible explanation for the experimental results.[18]

Falk and Kosfeld themselves argue verbally that the results are driven by what they call distrust aversion, i.e. that an agent dislikes being distrusted by another player and responds by choosing lower effort levels when a restriction is imposed. However, note that the experiment was anonymous, i.e. principals and agents did not meet. Hence, a principal choosing a restriction did not distrust this agent in person as agents were randomly assigned and she did not know with which agent she was playing. Hence, the choice of a restriction rather conveyed a principal's trust in the distribution of types among all participants in the experiment and not her trust in the particular agent, which is well in line with our approach.[19]

5 Trust and Employee Self-Selection

So far we have studied a model in which an employer's decision on whether to trust or to control her employees only has an impact on the moral convictions of a given set of employees. But of course such a decision will also affect the attractiveness of the job and, hence, the selection of agents working for the firm. To analyze such selection effects of trust in a simple way, we extend our framework by assuming that the principal initially employs a continuum of agents indexed by $i \in [0, 1]$, drawn from some larger population. After having received a signal about the fraction of fair agents among the steadfasts the principal decides on whether to trust or to control the agents.

As before, the agents then update their beliefs about ϕ based on the principal's decision. But now we add an additional stage at which the agents receive outside offers and leave the firm when these offers yield higher utility levels than staying with the firm. The utility level u_i generated by these offers is a random variable characterized by a cumulative distribution function $F_S(u_i)$ for a selfish and $F_F(u_i)$ for a fair type. Conformists fall into these

[18] The result can also be generalized by allowing for the possibility that principal and agent receive different but equally precise signals about the fraction of fair agents in advance.

[19] Recently, Ellingsen and Johannesson (2005) also gave a theoretical explanation for the experimental results. Their explanation is driven by agents' preferences for the principal's esteem which also seems problematic due to the anonymity of interaction in the experiment.

categories according to their beliefs on ϕ. For simplicity, we assume that when being trusted all agents prefer to stay with the organization.[20] When being controlled the continuation utility of a fair agent who stays with the firm is U_{FC} and that of a selfish agent is U_{SC}. As we want to investigate the costs of employing selfish agents in a simple way, we furthermore assume that the principal makes losses with any selfish agent working for the firm, i.e. $\Pi_{ST}, \Pi_{SC} < 0$.

Now, the decision whether to trust or to control employees not only affects conformists moral convictions but also the overall distribution of types in the organization. We investigate two cases. Either the conformists' convictions are affected by their beliefs about the distribution of types in the *whole population* or they consider only the type distribution among their colleagues *within the organization* they work for.

5.1 Population Norms

Here we assume that conformists are fair if and only if they believe that the median steadfast agent in the population is fair. We again check whether a separating equilibrium can exist in which trust is a credible signal of a favorable social norm. Suppose that this would be the case. The principal's expected profit when controlling ($\tau = C$) then becomes

$$((1-\eta)(1-\phi)+\eta) \cdot F_S(U_{SC}) \cdot \Pi_{SC} + (1-\eta)\phi \cdot F_F(U_{FC}) \cdot \Pi_{FC} \quad (3)$$

and when she trusts ($\tau = T$) she earns

$$(1-\eta)(1-\phi) \cdot \Pi_{ST} + (\eta + (1-\eta)\phi) \cdot \Pi_{FT}. \quad (4)$$

When the principal controls, selfish types leave the organization but fair types may also quit. Whereas the former effect is beneficial, the latter is costly for the principal. Hence, these selection effects create an additional trade-off which has to be taken into account by the principal.

We can again proceed as in subsection 4.1 by comparing these two expressions and solving for η. We obtain that the principal prefers to trust

[20] This assumption can be relaxed to allow for the possibility that agents also leave the firm when being trusted. The results are for instance robust when agents of both types prefer being trusted and the fraction of the selfish agents who stay when the principal trusts is at least as large as that of the fair agents, which seems reasonable as the selfish agents typically should benefit more from not being controlled.

when the fraction of conformists is larger than a cut-off value

$$\tilde{\eta}(\phi) = 1 - \frac{\Pi_{FT} - F_S(U_{SC})\Pi_{SC}}{\Pi_{FT} - \Pi_{ST} - \phi(F_S(U_{SC})\Pi_{SC} - F_F(U_{FC})\Pi_{FC} + \Pi_{FT} - \Pi_{ST})}. \tag{5}$$

First, note that the cut-off is decreasing in $F_S(U_{SC})$.[21] The stronger the selection effect for the selfish types (i.e. the smaller $F_S(U_{SC})$) the higher the cut-off value: control is indeed more attractive when it serves to make many selfish agents quit. But the cut-off is increasing in $F_F(U_{FC})$, i.e. trust becomes more advantageous when the selection effect for the fair types is stronger.

It is important to note that irrespective of which of the two effects dominates, trust is always beneficial when the fraction of conformists is sufficiently large – provided that it is a credible signal of a favorable social norm. In the following result we characterize under which conditions this will be the case:

Proposition 4 *If the fraction of fair agents staying with the firm when being controlled $F_F(U_{FC})$ is sufficiently large, that is if*

$$F_F(U_{FC}) > \frac{1}{\Pi_{FC}} \left(\Pi_{FT} - \frac{1-\phi_L}{\phi_L} \left(F_S(U_{SC}) \Pi_{SC} - \Pi_{ST} \right) \right), \tag{6}$$

a separating equilibrium exists in which the principal trusts after he has received the good signal and controls after the bad if and only if the fraction of conformists $\eta \in [\max\{0, \tilde{\eta}(\phi_H)\}, \tilde{\eta}(\phi_L)]$, where $\tilde{\eta}(\phi_H) < \tilde{\eta}(\phi_L) < 1$ and $\tilde{\eta}(\phi_L) > 0$.

Proof: See Appendix.

Without selection effects the only motive to trust the agents was to signal a favorable work norm. Now trust can also become beneficial without the signaling motive if control drives away too many of the fair agents. Condition (6) is equivalent to the requirement that $\tilde{\eta}(\phi_L) > 0$. If the condition is not met, the selection effect for the fair types is very strong relative to that for the selfish types. In that case, the principal will always choose to trust irrespective of the proportion of conformists and his signal. But in turn, trust is then no longer a credible signal of a favorable work norm.

[21] To see this, note that $\tilde{\eta}(\phi) = \frac{\phi(F_F(U_{FC})\Pi_{FC} - \Pi_{FT}) + (1-\phi)(F_S(U_{SC})\Pi_{SC} - \Pi_{ST})}{\phi(F_F(U_{FC})\Pi_{FC} - F_S(U_{SC})\Pi_{SC}) + (1-\phi)(\Pi_{FT} - \Pi_{ST})}$.

Condition (6) for instance always holds when $F_F(U_{FC}) = 1$, i.e. when all fair types stay with the firm when being controlled. This is for instance reasonable for the application on imposing restrictions considered in subsection 4.3. In that case fair agents are indifferent between being controlled or trusted as they choose the same effort level in both cases. The selection effect then only drives away the selfish types, which of course makes control attractive. But as the result shows, the separating equilibrium still exists in which an optimistic principal credibly signals trust if the fraction of conformists is neither too large nor too small. Of course, the selection effect increases the opportunity costs of trust and hence the signaling costs. On the one hand, this makes signaling harder as trust is less attractive for an optimistic principal. But it also makes it easier, as trusting becomes less attractive for a pessimistic principal, which strengthens the credibility of the signal.

5.2 Organization-Specific Norms

It is also interesting to investigate the case where conformists follow the social preferences only of their fellow employees within the particular organization they work for. To study this we now assume that conformists become fair when they believe that the median fellow employee in the considered organization is fair. Note that the decision on whether to trust or to control now affects the conformists' preferences in two ways. As before, there may be a signaling effect as the principal's superior information on the distribution of types may be revealed. But in addition, there is now also a direct effect on the social norm as the decision alters the composition of the workforce.

Note that the distribution of types in the organization remains unchanged when the principal trusts her employees. But when she controls them the proportion of fair agents among the steadfasts in the organization becomes

$$\frac{\phi F_F(U_{FC})}{\phi F_F(U_{FC}) + (1-\phi) F_S(U_{SC})}. \tag{7}$$

This proportion is larger than the population share ϕ whenever $F_F(U_{FC}) > F_S(U_{SC})$, i.e. when control turns away more selfish agents than fair ones. As it seems reasonable that selfish agents suffer more from being controlled, we assume that this is indeed the case.

As before, we check whether a separating equilibrium exists in which trusts is a credible signal of a favorable norm. If this is the case the conformists will again be fair when being trusted. But when being controlled they now become selfish only if for $\phi = \phi_L$ expression (7) is smaller than a half, which is equivalent to

$$\frac{\phi_L}{1-\phi_L} F_F\left(U_{FC}\right) < F_S\left(U_{SC}\right). \tag{8}$$

When condition (8) holds, the organization-specific norm is identical to the population norm and we can directly apply the results from proposition 4 and show that trust may again be a credible signal.

But this is not necessarily the case. Condition (8) does not hold when the selection effect for the selfish agents is very strong relative to that for the fair agents. In this case control actually leads the conformists to become fair. The principal's decision then becomes a pure selection decision as the conformists are fair irrespective of her choice. When for instance $F_F\left(U_{FC}\right) = 1$ such that fair agents always stay with the organization it is straightforward to see that the principal would always prefer to control[22] and the separating equilibrium can no longer exist.

The reason for this result is that here control repels so many selfish agents that the fair agents constitute a majority within the organization even when the share of steadfastly fair agents in the initial population is small. This effect hints at a possible benefit of control in organizations: when conformist employees follow the behavior of their direct colleagues rather than being influenced by the ethics of society as a whole, control may strengthen the work norm in the organization. An important precondition for this effect is of course that employees with a high work ethic do not care about being controlled but control drives most selfish employees away. When this is not the case, trust can still be beneficial as a credible signal of a favorable work norm.

[22] Note that when $F_F\left(U_{FC}\right) < 1$ trust may still be beneficial even without a signaling effect. This is the case when Π_{FT} is very large and only weakly smaller than Π_{FC} such that losing even only a few fair agents is very costly.

6 Further Analysis and Extensions[23]

6.1 Pooling Equilibria

We now return to the initial set-up and investigate all other feasible pure-strategy equilibria. It is straightforward that there is no separating equilibrium in which an optimistic principal distrusts and a pessimistic principal trusts. In a pooling equilibrium agents do not learn anything from the principal's choice. The conformists' behavior then depends on the common prior belief about the social norm.

When according to public information the norm is to be trustworthy ($E[\phi] \geq \bar{\phi}$) the conformists will remain trustworthy in equilibrium. In this case there will always be a pooling equilibrium in which the principal controls whatever her private information as control will not be perceived as a signal of pessimism about the social norm. But when the fraction of conformists is very large, another pooling equilibrium exists in which the principal always trusts. To see the latter, note that control is in this case a deviation from the equilibrium path. Such an equilibrium can be sustained if the agents believe after a deviation that the principal is pessimistic, and if even a pessimistic principal then prefers to trust which is the case when $\eta \geq \hat{\eta}(\phi_L)$. We can conclude:[24]

Proposition 5 *When $E[\phi] \geq \bar{\phi}$ (i) there is always a pooling equilibrium in which the principal controls whatever her signal. (ii) A pooling equilibrium in which she always trusts exists if and only if the fraction of conformists is sufficiently large, i.e. $\eta \geq \hat{\eta}(\phi_L)$ and $\Pi_{FT} \geq \Pi_{SC}$.*

But when according to the public information most steadfast agents are selfish ($E[\phi] < \bar{\phi}$), always trusting cannot be an optimal strategy as trust is then no signal of optimism about the norm. However, neither is always controlling an equilibrium strategy in many cases when we require intuitively plausible beliefs off the equilibrium path as the following result shows :

Proposition 6 *When $E[\phi] < \bar{\phi}$ (i) there is never a pooling equilibrium in which the principal always trusts. (ii) Given that the intuitive criterion is*

[23] This section contains extensions which are not part of the published version of the paper.

[24] It can be easily checked that both equilibria satisfy the Cho-Kreps (1987) Intuitive Criterion.

applied, a pooling equilibrium in which the principal always controls exists if and only if the fraction of conformists $\eta \notin [\hat{\eta}(\phi_H), \hat{\eta}(\phi_L)]$.

Proof: See Appendix.

The key intuition for this result is as follows: Suppose there would be a pooling equilibrium in which the principal always controls. An optimistic principal would prefer to deviate and trust when this makes conformists trustworthy if the fraction of conformists is not too small (i.e. when $\eta \geq \hat{\eta}(\phi_H)$). But when this fraction is not too large (i.e. when $\eta \leq \hat{\eta}(\phi_L)$), a pessimistic principal is always better off sticking to the equilibrium strategy. Hence, it is implausible to believe that the principal received the bad signal when she trusts and this makes the deviation for the optimistic principal attractive. An important consequence of this result is that the separating equilibrium we analyzed in Proposition 1 is the unique pure-strategy equilibrium in this case.

6.2 The Principal's Prior

So far we assumed that the principal has superior information about ϕ. To endogenize this, consider a model in which the principal employs a continuum of agents indexed by $i \in [0,1]$ in two periods $t = 1, 2$. Now we assume that principal and agents initially do not know the fraction of fair agents among the steadfasts ϕ which is either ϕ_H or ϕ_L with given probabilities where $\phi_H > \phi_L$. Hence, in contrast to the basic model principal and agents are symmetrically informed at the outset. Furthermore, we assume that all know the fraction η of conformists.

At the beginning of each period t the principal decides whether to trust or to control the agents $\tau_t \in \{T, C\}$ and each agent chooses his effort level $e_t(i)$ leading to a profit for the principal of $\Pi_t(i)$. Only the principal observes the aggregate performance of all agents after each period

$$\Pi_t = \int_0^1 \Pi_t(i)\, di.$$

The atomistic agents do not have a strategic motive in period 1 as an individual agent's effort choice has no impact on the principal's second-period decision. As before, profits from individual agents therefore take only two values depending on the principal's decision τ_t, either $\Pi_{F\tau_t}$ or $\Pi_{S\tau_t}$. The

profit generated by a conformist, which we denote by Π_C, is either $\Pi_{F\tau_t}$ or $\Pi_{S\tau_t}$ depending on the prior beliefs about ϕ. But then the principal can infer the behavior of the conformists. Hence, after period 1 she perfectly learns ϕ from observing aggregate profits as

$$\Pi_t = \eta\Pi_C + (1-\eta)\left(\phi\Pi_{F\tau_t} + (1-\phi)\,\Pi_{S\tau_t}\right) \Leftrightarrow \phi = \frac{\frac{\Pi_t - \eta\Pi_C}{1-\eta} - \Pi_{S\tau_t}}{\Pi_{F\tau_t} - \Pi_{S\tau_t}}.$$

Each agent knows that the principal can infer ϕ after the first period and therefore is aware of her superior information in the second. But then the second-period equilibrium analysis proceeds exactly as in the single-agent model above, where the probabilities in equations (2) and (1) now denote the mass of steadfastly fair, steadfastly selfish and conformist agents respectively.

6.3 Optimal Incentive Schemes

In subsection 4.2 we assumed that the principal chooses between two given contracts. Now we show that the key result carries over to a case where she can choose a contract $(w, \beta) \in \mathbb{R}^2$ and the agent responds by exerting effort level e.[25] The agent is protected by limited liability. For simplicity we therefore assume that his monetary income must always exceed a reservation income normalized to 0.

A Perfect Bayesian Equilibrium is characterized by a pair of contracts (w_s, β_s) for each of the principal's signals $(s = L, H)$, the conformist's beliefs about ϕ for each possible contract choice, and the different types' effort choices given their beliefs and the offered contract. Suppose now that a separating equilibrium exists in which the principal chooses two different contracts depending on her signal. When offering (w_L, β_L) the agents infer that she has received the bad signal and conformists will act as outsiders. The expected fraction of insiders is then $(1-\eta)\,\phi_L$ and the contract offered by a pessimistic principal is uniquely determined: it consists of a fixed wage

[25] Note that we do not investigate screening here as only a single contract can be offered. A justifaction for this is that typically, firms offer the same type of contract to employees who perform the same function. But even with screening a signaling effect may arise as the optimal menu will depend on the principal's beliefs about the type distribution.

of $w_L = 0$ and a piece rate β_L solving

$$\begin{aligned} \max_{\beta_L} \quad & \left(\left(1-\eta\right)\phi_L e_F + \left(1-\left(1-\eta\right)\phi_L\right)e_S\right)\left(1-\beta_L\right) \\ \text{s.t.} \quad & e_S = \frac{\beta_L}{c} \text{ and } e_F = \frac{\left(1-\mu\right)\beta_L + \mu}{c} \end{aligned}$$

which is equal to

$$\beta_L = \frac{1-2\left(1-\eta\right)\phi_L\mu}{2-2\left(1-\eta\right)\phi_L\mu}.$$

But using this β_L, the conditions derived in Proposition 2 also characterize all fixed-wage contracts offered by an optimistic principal that can be sustained in a separating equilibrium.[26] To see this note the following: If a certain fixed wage contract can be sustained in equilibrium, it will always be sustainable with the worst possible beliefs of the agent after any deviation off the equilibrium path, i.e. when agents then believe that the principal has received the bad signal. Hence, contract $(w_H, 0)$ is part of a separating equilibrium whenever (i) an optimistic principal's is better off with this contract when conformists become insiders than with $(0, \beta_L)$ when they remain outsiders as all other feasible deviations are dominated by the latter and (ii) a pessimistic principal prefers the incentive scheme $(0, \beta_L)$ to the fixed-wage contract. But this is exactly what is checked in Proposition 2.

7 Conclusion

From a more general perspective our model may yield some insights on the notion of trust. Trust can be straightforwardly defined in social preference frameworks: *trust in a transaction partner is the belief that this transaction partner has social preferences with a sufficiently high probability instead of being selfish.* But as experiments have shown, trust seems to affect the trustworthiness of the transaction partner, which cannot be explained by distributional theories of social preferences alone.

Our model suggests the following mechanism for this phenomenon: not trusting a person reveals your belief that there is a danger that this person is selfish and will choose a harmful action. A reason for distrusting someone is that you have had a bad experience in a similar situation before and therefore you are pessimistic about the trustworthiness of your counterpart.

[26] Just set $w = 0$ and $\Delta = w_H$.

Distrust hence reveals your belief about the typical behavior in a reference group. If your counterpart is influenced by social norms, i.e. his beliefs about what others would do in the same situation, this information may then indeed let him become selfish. On the other hand, trusting a person may reveal your confidence that a person with such characteristics would not be selfish. But this in turn makes being selfish more "costly" for the person as it reveals that not being selfish is the social norm. It is exactly the danger of being harmed that makes signaling possible. The danger is lower for people who are confident that their counterparts "can be trusted".

Note that this explanation for motivation crowding-out is distinct from those proposed by psychologists who have mainly focused on changes in the enjoyment of a task.[27] Our explanation also works when agents perfectly know whether they like or dislike a task itself such as in typical principal agent models. It of course rests on the assumption that there is some uncertainty about norms of behavior. This should be especially relevant in larger firms where employees cannot perfectly observe the behavior of all others working on similar tasks or in newly formed departments. However, it should be less relevant in firms where all employees can mutually observe their respective efforts. Hence, our model may yield some indication for why crowding-out has so often been observed in experiments where the situation is always new to the participants and there is always uncertainty about the "appropriate" behavior.

In the model we applied a notion of social norms where the norm does not specify a particular *action* but a more general rule of behavior forming the *intentions behind* the chosen action. However, there is also the different view that social norms define specific actions chosen by individuals in a reference group.[28] We can model this in our framework in a very simple

[27] Very roughly, cognitive evaluation theory (see for instance Deci and Porac (1978)) posits that monetary rewards undermine self-determination and therefore the joy of performing a task. According to self-perception theories individuals imperfectly know their preferences and incentives may lead individuals to conclude that they perform an activity because of those incentives (see Lepper and Greene (1978)).

[28] Hence, in the first approach the norm is "to be fair" whereas in the second it requires to choose a particular effort level. Recent experimental evidence (see for instance Charness and Rabin (2002), Falk et al. (2000)) suggests that learning about other people's intentions indeed affects decisions even if the material consequences remain unchanged.

The intention or type-based approach is related to Bernheim (1994)'s model of conformism where individuals care for others' actions only indirectly as these actions reveal something about their type. The action-based approach is closer to peer pressure in team models such as Kandel and Lazear (1992).

way by assuming that a conformist's utility is always maximized when his effort level is equal to the effort level chosen by the median agent. When conformists consider only the actions of the steadfast agents, a separating equilibrium then exists under the same conditions as given in Proposition 1.

When conformists consider the whole population (including their fellow conformists) there are typically multiple equilibria at the effort stage and different norms can be stable. The conformists play a coordination game if neither the steadfastly fair nor the steadfastly selfish agents constitute a majority. However, a simple refinement again establishes the separating equilibrium laid out above: suppose that the expected behavior of the median steadfast agent forms a focal point and the equilibrium effort level is selected accordingly. Then the best response of each individual conformist will be to put in the fair effort when he is trusted and the selfish effort level when he is controlled. The mass of agents choosing the respective action will indeed always form a majority. Hence, an employer's choice of whether to trust or to control may guide the coordination on a work norm in a team.

8 Appendix

Proof of Proposition 1:
As $\phi_H > \frac{1}{2} > \phi_L$ the conformists will indeed be trustworthy in a separating equilibrium if and only if the principal trusts. Hence, such an equilibrium will exist if she only prefers to trust after the high signal. This will be the case whenever $\eta \in [\hat{\eta}(\phi_H), \hat{\eta}(\phi_L)]$. Note that $\hat{\eta}(\phi_L) > \hat{\eta}(\phi_H)$ and from properties 1 and 3 we can conclude that

$$0 < \hat{\eta}(1) = 1 - \frac{\Pi_{FT} - \Pi_{SC}}{\Pi_{FC} - \Pi_{SC}} < \hat{\eta}(0) = 1 - \frac{\Pi_{FT} - \Pi_{SC}}{\Pi_{FT} - \Pi_{ST}} \leq 1.$$

when $\Pi_{FT} \geq \Pi_{SC}$. Therefore the set is non-empty. ■

Proof of Proposition 2:
We know that for a given ϕ the principal will trust iff $\eta \geq \hat{\eta}(\phi)$, where now

$$\hat{\eta}(\phi) = 1 - \frac{\mu - \beta(1-\beta) - \Delta c}{\mu(1 - \phi\beta(2-\beta))}.$$

the denominator of the second term is always positive as $\max_\beta \beta(2-\beta) = 1$. Hence, this condition is equivalent to

$$\Delta \leq \mu \frac{\eta + (1-\eta)(2-\beta)\beta\phi - \beta(1-\beta)}{c}.$$

Therefore the separating equilibrium exists iff Δ is in the given interval. Both boundaries intersect at $\frac{\mu-\beta(1-\beta)}{c}$ when $\eta = 1$. Each is linearly increasing in η and the slope of each is larger, the smaller ϕ. Hence, the set is non-empty as long as the upper boundary exceeds 0, which is always the case when η is large enough and $\mu \geq \beta(1-\beta)$. ■

Proof of Proposition 3:
The first part follows directly from Proposition 1. The expected effort level without restriction is $(\eta + (1-\eta)\phi_H)e^*$ in the separating equilibrium. With a restriction it is $(1-(1-\eta)\phi_L)R + (1-\eta)\phi_L e^*$. But

$$\begin{aligned}(\eta + (1-\eta)\phi_H)e^* &\geq (1-(1-\eta)\phi_H)R + (1-\eta)\phi_H e^* \\ &\geq (1-(1-\eta)\phi_L)R + (1-\eta)\phi_L e^*\end{aligned}$$

where the first inequality follows as the optimistic principal prefers to trust in the separating equilibrium. ■

Proof of Proposition 4:
This cut-off $\tilde{\eta}(\phi)$ given by (5) is strictly decreasing in ϕ, when $F_S(U_{SC})\Pi_{SC} - \Pi_{ST} > F_F(U_{FC})\Pi_{FC} - \Pi_{FT}$. As

$$F_S(U_{SC})\Pi_{SC} - \Pi_{ST} > \Pi_{SC} - \Pi_{ST} > \Pi_{FC} - \Pi_{FT} > F_F(U_{FC})\Pi_{FC} - \Pi_{FT}$$

this is always the case. Hence, the separating equilibrium will exist whenever $\eta \in [\tilde{\eta}(\phi_H), \tilde{\eta}(\phi_L)]$. The cut-off value $\tilde{\eta}(\phi)$ is always smaller than one as

both numerator and denominator of

$$\frac{\Pi_{FT} - F_S(U_{SC})\Pi_{SC}}{\Pi_{FT} - \Pi_{ST} - \phi(F_S(U_{SC})\Pi_{SC} - F_F(U_{FC})\Pi_{FC} + \Pi_{FT} - \Pi_{ST})}$$

are strictly positive. $\tilde{\eta}(\phi)$ is strictly positive for a given ϕ when

$$\begin{aligned}
&\Pi_{FT} - \Pi_{ST} - \phi(F_S(U_{SC})\Pi_{SC} - F_F(U_{FC})\Pi_{FC} + \Pi_{FT} - \Pi_{ST}) \\
&\qquad > \Pi_{FT} - F_S(U_{SC})\Pi_{SC} \\
&\Leftrightarrow F_F(U_{FC}) > \tfrac{1}{\Pi_{FC}}\left(\Pi_{FT} - \tfrac{1-\phi}{\phi}(F_S(U_{SC})\Pi_{SC} - \Pi_{ST})\right)
\end{aligned}$$

Note that the right hand side is smaller than 1 when

$$\tfrac{1-\phi}{\phi}(\Pi_{ST} - F_S(U_{SC})\Pi_{SC}) < \Pi_{FC} - \Pi_{FT}$$

which always holds as $\Pi_{ST} < F_S(U_{SC})\Pi_{SC} < 0$. ■

Proof of Proposition 6:
A pooling equilibrium where the principal always controls exists when the agent believes off the equilibrium path that the principal received the low signal. But these beliefs may be ruled out by the intuitive criterion: when the principal received ϕ_L her equilibrium payoff is

$$((1-\eta)(1-\phi_L) + \eta)\Pi_{SC} + (1-\eta)\phi_L\Pi_{FC}$$

as all conformists are selfish. However, the highest possible payoff when deviating to trust is

$$(1-\eta)(1-\phi_L)\Pi_{ST} + ((1-\eta)\phi_L + \eta)\Pi_{FT}.$$

The equilibrium payoff after ϕ_L exceeds the highest feasible payoff when deviating, whenever $\eta \leq \hat{\eta}(\phi_L)$. Hence, in this case a deviation is equilibrium-dominated for a pessimistic principal and will then lead the agents to believe in a high signal. An optimistic principal will therefore choose trust whenever $\eta \geq \hat{\eta}(\phi_H)$. ■

References

Akerlof, G. (1980): A Theory of Social Custom, of Which Unemployment May be One Consequence. Quarterly Journal of Economics, 94, pp. 749–75.

Akerlof, G. and Kranton, R. (2005): Identity and the Economics of Organizations. Journal of Economic Perspectives, 19, pp. 9–32.

Allen, F. and Gale, D. (1992): Measurement Distortion and Missing Contingencies in Optimal Contracts. Economic Theory, 2, pp. 1–26.

Baron, J. N. and Kreps, D. M. (1999): Strategic Human Resources: Frameworks for General Managers. Wiley, New York.

Bernheim, B. D. and Whinston, M. D. (1998): Incomplete Contracts and Strategic Ambiguity. American Economic Review, 88, pp. 902–32.

Bernheim, D. B. (1994): A Theory of Conformity. Journal of Political Economy, 100, pp. 841–7.

Bolton, G. and Ockenfels, A. (2000): ERC - A Theory of Equity, Reciprocity and Competition. American Economic Review, 90, pp. 166–193.

Bénabou, R. and Tirole, J. (2003): Intrinsic and Extrinsic Motivation. Review of Economic Studies, 70, pp. 489–520.

Charness, G. and Rabin, M. (2002): Understanding Social Preferences with Simple Tests. Quarterly Journal of Economics, 117, pp. 817–869.

Cho, I. and Kreps, D. (1987): Signaling Games and Stable Equilibria. Quarterly Journal of Economics, 102, pp. 179–222.

Clark, A. (2003): Unemployment as a Social Norm: Psychological Evidence from Panel Data. Journal of Labor Economics, 21, pp. 323–351.

Deci, E. L. (1971): Effects of Externally Mediated Rewards on Intrinsic Motivation. Journal of Personality and Social Psychology, 18, pp. 105–115.

Deci, E. L. and Porac, J. (1978): Cognitive Evaluation Theory and the Study of Human Motivation. In: Lepper, M. R. and Greene, D. (Ed.) The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation, Lawrence Erlbaum, Hillsdale, NJ, pp. 149–175.

Ellingsen, T. and Johannesson, M. (2005): Trust as an Incentive. Mimeo Stockholm School of Economics.

Falk, A., Fehr, E. and Fischbacher, U. (2000): Testing Theories of Fairness - Intentions Matter. Working Paper No. 63 University of Zurich.

Falk, A. and Kosfeld, M. (forthcoming): Distrust - The Hidden Cost of Control. American Economic Review.

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

Fehr, E. and Rockenbach, B. (2003): Detrimental effects of sanctions on human altruism. Nature, 422, pp. 137–140.

Fehr, E. and Schmidt, K. M. (1999): A Theory of Fairness, Competition, and Cooperation. Quarterly Journal of Economics, 114, pp. 817–868.

Fischer, P. E. and Huddart, S. J. (2005): Optimal Contracting with Endogenous Social Norms. SSRN Working Paper.

Frey, B. S. (1997): Not Just For the Money. An Economic Theory of Personal Motivation. Edward Elgar, Cheltenham.

Frey, B. S. and Jegen, R. (2001): Motivation Crowding Theory: A Survey of Empirical Evidence. Journal of Economic Surveys, 15 (5), pp. 589–611.

Frey, B. S. and Oberholzer-Gee, F. (1997): The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out. American Economic Review, 87, pp. 746–755.

Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989): Psychological Games and Sequential Rationality. Games and Economic Behavior, 1, pp. 60–79.

Gibbons, R. S. (1997): Incentives and Careers in Organizations. In: Kreps, D. M. and Wallis, K. (Ed.) Advances in Economic Theory and Econometric: 7th World Congress of the Econometric Society, Cambridge University Press, Cambridge.

Gneezy, U. and Rustichini, A. (2000a): Pay Enough or Don't Pay at All. Quarterly Journal of Economics, 115, pp. 791–810.

Holmström, B. and Milgrom, P. (1991): Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design. Journal of Law, Economics and Organization, 7, pp. 24–52.

Huck, S., Kübler, D. and Weibull, J. (2003): Social Norms and Economic Incentives in Firms. ELSE Working Paper.

Ichino, A. and Maggi, G. (2000): Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm. Quarterly Journal of Economics, 115, pp. 1057–1090.

Irlenbusch, B. and Sliwka, D. (2003): Incentives, Decision Frames and Motivation Crowding Out - An Experimental Investigation. Mimeo, Universität Bonn.

Kandel, E. and Lazear, E. P. (1992): Peer Pressure and Partnerships. Journal of Political Economy, 100, pp. 801–17.

Kreps, D. M. (1997): Intrinsic Motivation and Extrinsic Incentives. American Economic Review, 87 (2), pp. 359–64.

Kunz, A. and Pfaff, D. (2002): Agency Theory, Performance Evaluation, and the Hypothetical Construct of Intrinsic Motivation. Accounting, Economics and Society, 27, pp. 275–295.

Lazear, E. P. (2000): Performance Pay and Productivity. American Economic Review, 90 (5), pp. 1346–62.

Lepper, M. R. and Greene, D. (1978): Overjustification Research and Beyond: Towards a Means-Ends Analysis of Intrinsic and Extrinsic Motivation. In: Lepper, M. R. and Greene, D. (Ed.) The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation, Lawrence Erlbaum, Hillsdale, NJ, pp. 109–148.

Levine, D. (1998): Modeling Altruism and Spitefulness in Experiments. Review of Economic Dynamics, 1, pp. 593–622.

Lindbeck, A., Nyberg, S. and Weibull, J. (1999): Social Norms and Economic Incentives in the Welfare State. Quarterly Journal of Economics, 114, pp. 1–35.

Moffitt, R. (1983): An Economic Model of Welfare Stigma. American Economic Review, 73, pp. 1023–1035.

Nagin, D. S., Rebitzer, J. B., Sanders, S. and Taylor, L. J. (2002): Monitoring, Motivation and Management: The Determinants of Opportunistic Behavior in a Field Experiment. American Economic Review, 92, pp. 850–873.

Parent, D. (2002): Incentive Pay in the United States: Its Determinants and Its Effects. In: Brown, M. and Heywood, J. (Ed.) Paying for Performance: An International Comparison, Sharpe, pp. 17–51.

Parent, D. and MacLeod, W. B. (1999): Job Characteristics and the Form of Compensation. Research in Labor Economics, 18.

Prendergast, C. J. (1999): The Provision of Incentives in Firms. Journal of Economic Literature, 37, pp. 7–63.

Rabin, M. (1993): Incorporating Fairness Into Game Theory and Economics. American Economic Review, 83, pp. 1281–1302.

Spier, K. E. (1992): Incomplete Contracts and Signaling. Rand Journal of Economics, 23, pp. 432–443.

Stutzer, A. and Lalive, R. (2004): The Role of Social Work Norms in Job Searching and Subjective Well-Being. Journal of the European Economic Association, 2, pp. 696–719.