TSE working paper series

09-078

Research Group: *Behavioral and Experimental Economics*

August, 2009

Contracting under Reciprocal Altruism

OLEG SHCHETININ

Ecole d'économie de Toulouse (TSE) - Manufacture des Tabacs
Aile Jean-Jacques Laffont - 21, allée de Brienne - 31000 TOULOUSE
Tél : +33 (0) 5 61 12 85 89 - Fax : +33 (0) 5 61 12 86 37 - **www.TSE-fr.eu** - contact@TSE-fr.eu

Contracting under Reciprocal Altruism *

Oleg Shchetinin

Toulouse School of Economics

August 2009

Abstract

I develop a model of reciprocal altruism which accounts for some evidence in contracting situations, which are paradoxical from the point of view of neoclassical contract theory with selfish actors. My model predicts the crowding-out effect, observed in the Trust Game with the possibility of a fine; for the Control Game the model predicts that an equilibrium can exhibit "no effect of control", "hidden cost of control", or "positive effect of control", depending on the characteristics of the actors, as observed in the experiments. This suggests that reciprocal altruism modeling could be fruitful more generally in applications of contract theory.

Keywords: Contract Theory, Signaling, Behavioral Economics, Reciprocal Altruism, Extrinsic and intrinsic motivation, Experimental Economics.

JEL Classification Numbers: D82, M54

1 Introduction

Intriguing observations about human response to incentives have recently been made. For instance, providing additional incentives can, in contrast with standard models with selfish actors, lead to lower levels of performance and intentions seem to matter, according to Fehr and Rockenbach [2003], Falk and Kosfeld [2006] and many others[1].

In this paper I develop a Principal-Agent model embodying reciprocal altruism. The paper shows that a simple formal model of reciprocal altruism is able to give reliable predictions for some patterns of human behavior, puzzling when considered within the standard selfish paradigm. While the idea that reciprocity, altruism and other forms of social preferences shape people's behavior is not new, there are only a few models of reciprocal altruism in the literature.

My model is based on the premise that a person cares more about those who care more about him. In other words, a person is more altruistic towards

*I'm grateful to Jean Tirole for many insightful discussions. I'm also grateful to Stefanie Brillon, Andrey Bremzen, Robert Dur, Arjan Non, the participants of the BEE workshop in Toulouse, and the participants at seminars at ECARES(Brussels), Gothenburg University, CEFIR(Moscow) for their comments. All errors are mine.

[1]See below the more detailed discussion of the relevant literature.

those whom he perceives as being altruistic towards him. This is the essence of the reciprocal altruism. In a Principal-Agent relationship, an altruistic Agent is inspired to exert effort even in the absence of monetary incentives, i.e. the Agent's altruism works as an intrinsic motivator. If furthermore, the Agent is reciprocal, the Principal will want to demonstrate his altruism in order to boost the Agent's intrinsic motivation. This leads to a signaling game in which the Principal signals his altruism through offering a "generous" contract.

I assume that the population of the Principals and the Agents is heterogenous: together with selfish actors there are pro-social ones, who are more altruistic and reciprocal.

I consider two variants of the model, closely related to existing lab experiments, and so the model's equilibrium can be tested by the experiments' outcomes.

In the first setting, the Principal can either restrict the Agent's choice of effort by imposing minimal effort or give the Agent full flexibility, so that he can exert zero effort. Such contract resembles the Control Game experiment of Falk and Kosfeld [2006].

The predictions, based on the selfishness framework, are in striking contrast with what has been observed in the lab. In fact, if all the actors are selfish, in equilibrium the Principal restricts the Agent's effort, who, in turn, chooses the minimal feasible effort. Were the Agent uncontrolled, he would exert zero effort.

However, in the experiment the Principals often choose not to control and, after this, many Agents perform at much higher level than zero. Moreover, the average effort of the uncontrolled Agents is higher than that of the controlled ones for some values of the minimum effort cut-off. In other words, the Principal's action is often a deviation from the equilibrium path for selfish players, and after this deviation, many Agents deviate from the continuation subgame optimal move. I argue, however, that these "deviations" fit the equilibrium path if the actors's population has a pro-social component and is heterogenous.

Intuitively, the reciprocal (pro-social) Agents' intrinsic motivation is raised and so they perform at high level when they've learned that the Principal is pro-social, or generous. The Principal can signal her generosity through not restricting the Agent, i.e. by offering the generous contract. This explains why the pro-social Principals don't control on the equilibrium path and why the pro-social Agents respond by high performance. However, the selfish Agents' intrinsic motivation can't be boosted (because of weak or zero pro-social component in preferences) and they perform at zero level if not controlled. So, the observed performance of the uncontrolled Agents is either high or zero. Finally, the selfish Principals prefer to control and guarantee from all Agents a relatively low performance, equal to the controlling threshold, because they are less confident in the pro-sociality of the Agents' population than the pro-social Principals.

In the second setting, the Principal chooses whether to punish the Agent for low performance, and chooses the threshold below which the performance is considered to be low. After receiving the contract from the Principal, the Agent can accept or reject it. The contract resembles the variant of the Trust Game experiment of Fehr and Rockenbach [2003]. As in the first setting, the outcome of the experiment is puzzling if the actors are assumed to be selfish. In fact, within the selfishness framework, in equilibrium the Principal requires effort, while satisfying the Agent's participation constraint, and threatens the Agent with the punishment, which provides him an extrinsic incentive. If the Principal chooses not to punish, the Agent should perform at zero level, since extrinsic incentives aren't provided. However, in the experiment the Principals often choose not to punish, and, responding to this, many Agents choose to perform at the level which is much higher than zero. These behaviors clearly represent a deviation from the equilibrium path in the game with selfish actors. By contrast, when threatened with the punishment for low performance, most of the Agents choose the minimal performance level to avoid punishment, just as on the equilibrium path of the selfish players case.

As in the first setting, my analysis shows that the "deviating" actions lie on the equilibrium path in the signaling game with heterogenous actors and a pro-social component in the population. The mechanism behind the separating equilibrium with the same structure, as observed in the experiment, is similar to the one for the Control Game.

The behavioral patterns, supporting reciprocity and observed in the experiments are not limited to the lab, they are common in human relations, in the workplace in particular. The evidence from the field is documented in Gneezy [2002], Falk [2007], Bolton and Ockenfels [2008], Paarsch and Shearer [2007], Shearer [2004], Bellemare and Shearer [2007], Berry and Kanouse [1987], Maréchal and Thöni [2007].[2]

Broadly taken, my study contributes to the literature on behavioral theory of incentives, and one of its aims is explaining the puzzling behavior observed in the lab and in the field by taking into account the interaction between extrinsic incentives[3] and intrinsic motivation[4], which can lead, for instance, to the motivation crowding-out. I show that reciprocal altruism can account for many systematically observed behavior patterns given a natural information structure. Bénabou and Tirole [2003] and Bénabou and Tirole [2006a] argue that the impact of incentives on intrinsic motivation shapes behavior in many

[2] However, Kube et al. [2006] found support for negative reciprocity and question positive reciprocity, especially in the long-run. Gneezy and List [2006] found reciprocity in the short-run (the first 2 hours of work) and decreasing reciprocity in the long-run: to the end of the 6-hour job the subjects receiving a more generous wage didn't work harder than the others. Some studies question the relevance of the lab experiments - see, e.g. List [2007], Hennig-Schmidt et al. [2005], List and Levitt [2005]. We should be warned by these studies but evidence for reciprocity comes from many different sources, so it's hard to question that reciprocity is an important psychological characteristic of human beings.

[3] The list of extrinsic motivators is not limited to the incentive payments (piece-rate wage or bonus payment) but includes also expectation of future material payoff e.g. reputation building due to long-term interaction, strategic reciprocity, career concerns, comparative performance based payment (tournaments), monitoring/control etc.

[4] The literature provides evidence for many kinds of intrinsic motivation, apart from altruism and reciprocity. The Ultimatum Game introduced by Güth et al. [1982] illustrates that taste for fairness and/or inequality aversion is an important factor determining behavior; another evidence for fairness comes from different versions of the Gift Exchange Game - see Fehr et al. [1993], Fehr and Falk [1999]. Social norms (avoiding social disapproval/geting social approval) influence economic decisions. People can change their behavior under peer pressure or have a taste for the social embeddedness. The evidence are provided by a variant of the Gift Exchange Game in Gächter and Falk [2002] and Third Party Punishment Game by Fehr and Fischbacher [2004]. A person may have taste for the others' belief about his motivation (or type) - see Rabin [1993], Falk and Fischbacher [2006] and Bénabou and Tirole [2006a]. The list of intrinsic motivators can be continued with self-learning, working on interesting/challenging task (in this case effort may not be costly (painful), the job rather gives fun and higher effort increases utility) etc.

different contexts.

My paper is also related to the theoretical literature on reciprocity or social norms. Falk and Fischbacher [2006] develop a theory of reciprocity based on psychological games, i.e. with utilities of the actors depending not only on their material payoffs, but also on the perceived intentions of another player, i.e. on the 1-st and 2-nd order beliefs, following Rabin [1993]. The players' concern about the equitable outcome plays an important role in the analysis, in contrast with my model, which is based on reciprocal altruism. The model of Falk and Fischbacher [2006] can explain behavior in the Ultimatum Game, the Gift Exchange Game and some other experiments.

The model of Ellingsen and Johannesson [2008] can account for the crowding-out effect. It is similar to mine in that they consider altruistic actors. One difference is that they also incorporate the taste for social esteem (pride) in the utility function, i.e. second-order beliefs. Another difference is that the actors in their model are unconditionally altruistic.

The model of Sliwka [2007] can also account for crowding-out through a mechanism based on social norms, which is different from mine. There are reliable and unreliable Agents in his model. The reliable Agents follow the contract whereas the unreliable ones can deviate even if a contract is signed. As a consequence, the incentive scheme has to be high-powered for the unreliable Agents to perform at a high level. This leads to the fact that observing the high-powered incentive scheme, the Agent can learn that there is a social norm to be unreliable which can crowd-out his intrinsic motivation based on inherent reliability.

In a recent paper Dur [2008] analyzes a model based on reciprocal altruism, following Levine [1998] and putting it in the context of the workplace relation. He shows that non-contractible attention can substitute monetary compensation in a Principal-Agent relationship with reciprocal agent.

The paper proceeds as follows. Section 2 describes the framework for modeling reciprocal altruism and presents its general analysis, leading to the benchmark results. Section 3 applies the reciprocal altruism framework to develop models of interaction in two well-known lab experiments - the Trust Game with a possibility of a fine by Fehr and Rockenbach [2003] and the "Control Game" by Falk and Kosfeld [2006]. Section 4 develops a model of contracting with intrinsic incentives, based on reciprocal altruism only, without any extrinsic incentives and considers some applications of such contracting. Section 5 concludes.

2 The Reciprocal Altruism Framework

Consider a Principal-Agent relationship The Principal is altruistic towards the Agent and the Agent reciprocates the employer's altruism: if the Agent perceives the Principal to care about him, he becomes altruistic towards the Principal. The Principal offers a contract to the Agent.

Output is equal to effort, is observable and verifiable (can be contracted upon), so that there is no moral hazard.

Producing output is costly for the Agent. The cost function $C(q)$ satisfies standard assumptions - convexity and zero cost at zero output:

$$C'(q) > 0,\ C''(q) > 0 \text{ for } q > 0$$
$$C(0) = 0,\ C'(0) = 0$$

Let B be the Agent's exogenous benefit from interacting with the Principal[5]. The benefit can be psychological or a monetary payment from a third party[6].

For now, assume that the Agent doesn't respond to monetary incentives, beyond some subsistence level, that we normalize to zero. The selfish utilities of the Principal and the Agent are then given by

$$v = q$$
$$u = B - C(q)$$

Let α be the degree of the Principal's altruism and $\widehat{\alpha}$ denote the Agent's perception of the Principal's altruism. Let β denote the intensity of the Agent's reciprocity (more generally, it can be treated as intensity of intrinsic motivation of any nature emerging from perceiving the Principal as "generous"). The interaction term $\widehat{\alpha}\beta$ represents the Agent's altruism emerging as a result of reciprocating altruism of the Principal.

Assume that $\alpha \in [\alpha_1, \alpha_2] \subseteq [0, 1]$ and $\beta\alpha_2 \leq 1$. The assumptions mean that the Principal's and the Agent's altruism is less than 1, in other words the actors care about own material gain more than about the other's.

The utilities of the Principal and the Agent when the Agent produces output q are given by

$$V(q, \alpha) = v + \alpha u = q + \alpha(B - C(q)) \tag{1}$$
$$U(q, \widehat{\alpha}, \beta) = u + \widehat{\alpha}\beta v = B - C(q) + \widehat{\alpha}\beta q \tag{2}$$

The contract can be a command - "produce q" or can give the Agent some flexibility - say, "produce any quantity $q \in [q_1, q_2]$".

Notice the difference with the standard Principal-Agent setup. The Principal's valuation of the output is not always increasing, now it has an inverted-U shape: it increases only for small enough values of output and is maximal at some $q = q^P$. Similarly, the Agent's payoff is not always decreasing in output, and has an inverted-U shape: it decreases only for large enough values and reaches the maximal value at some q^A.

In what follows, I will refer to q^P and q^A as the Principal's and the Agent's preferred values of output (or performance). In contrast with the standard Principal-Agent models, $q^P \neq +\infty$, $q^A \neq 0$. Principal's and Agent's payoffs as functions of output are depicted in Figure 1.

For $\alpha = \beta = 1$ the Principal's and the Agent's interests are aligned, $U(q) \equiv V(q)$, because there is full internalization, so that the two curves representing the Principal's and the Agent's utilities in Figure 1 coincide.

For smaller values of α or β, i.e. weaker internalization, there is a conflict of interest like in the standard Principal-Agent setup but this conflict is softened

[5] More generally, B can be treated as opportunity cost of interacting with the Principal. In this case isn't necessarily positive.

[6] The latter is the case in the lab experiments which I consider in the paper. The third party will be an experimenter.

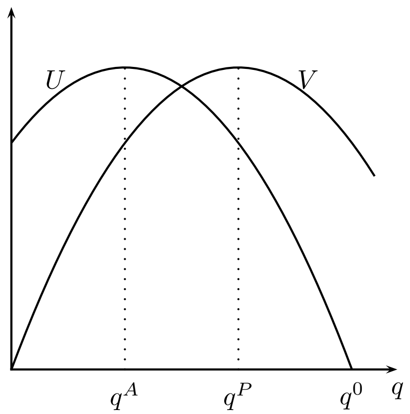

Figure 1: Principal's and Agent's payoffs under reciprocal altruism.

by the partial internalization of utilities. In the graph, the two inverted-U curves become more distant, and consequently, the distance between the maximizers of the Principal's and Agent's utilities q^P and q^A becomes larger: the Principal wants the Agent to exert more effort, whereas the Agent prefers performing less.

For an arbitrary $\widehat{\alpha}$ denote the value of q, making the participation constraint binding, by $q^0(\widehat{\alpha}\beta)$. I will refer to this value as the Agent's participation threshold.

For $\widehat{\alpha}\beta$ close to 1 the Agent's participation constraint is not binding because q^P is "close enough" to the maximizer of the Agent's utility q^A, where the Agent's utility is positive, and then, the Principal can implement her preferred output q^P. However, as α or $\widehat{\alpha}\beta$ decrease, the participation constraint becomes binding.

For specific applications of the reciprocal altruism framework, I will make additional assumption on the distributions of the Principal's and Agent's characteristics α and β and on the information structure. The details will be provided below.

2.1 Benchmark cases

The preferred output value for the Principal is given by

$$q^P(\alpha) = \arg\max_q [V(q,\alpha)] = \arg\max_q [q - \alpha C(q)]$$

or

$$C'(q^P) = \frac{1}{\alpha} \tag{3}$$

If there are no barriers to implementing this output level, such as Agent's participation constraint or limits on contract design, the Principal will induce it.

Lemma 1. *The Principal's preferred output $q^P(\alpha)$ is determined by* (3) *and is a decreasing function of α: $\frac{\partial q^P}{\partial \alpha} < 0$.*

The lemma follows directly from (3).

The preferred value of output for the Agent is given by

$$q^A(\widehat{\alpha}\beta) = \arg\max_q [U(q;\widehat{\alpha},\beta)] = \arg\max_q [\widehat{\alpha}\beta q - C(q)]$$

or

$$C'(q^A) = \widehat{\alpha}\beta \tag{4}$$

This output obtains when the Agent is given full flexibility or, more generally, if this level is available to the Agent despite some restrictions, such as binding contract, are imposed.

The Agent is willing to perform at an effort level such that his marginal cost is equal to his marginal benefit $\widehat{\alpha}\beta$. This means that $\widehat{\alpha}\beta$ is a measure of the Agent's intrinsic motivation, similarly to the monetary (extrinsic) incentives intensity.

Lemma 2. *The Agent's preferred output is determined by an increasing function* $q^A(\widehat{\alpha}\beta)$*, determined by* (4).

The lemma follows directly from (4).

For the case $\alpha < 1$ and $\widehat{\alpha}\beta < 1$ it's easy to see from (3) that $C'(q^P) > 1$, whereas (4) leads to $C'(q^A) < 1$, so that $q^P > q^A$ and there is always a gap between the Principal's and the Agent's preferred output levels. This gap is larger, the smaller α, $\widehat{\alpha}$ and β.

Lemma 3. *The Principal's preferred output is always larger than the Agent's preferred output, except when it is a common knowledge that* $\alpha = \beta = 1$*, in which case the preferred outputs are the same:*

$$q^P(\alpha) > q^A(\widehat{\alpha}\beta) \quad \textit{unless } \alpha = \widehat{\alpha} = 1, \qquad \beta = 1$$

$$q^P(1) = q^A(1)$$

For the case of symmetric or revealed information, so that $\widehat{\alpha} = \alpha$*, the gap* $q^P - q^A$ *between the Principal's and the Agent's preferred outputs is a decreasing function of* α, β*:* $\frac{\partial\left(q^P - q^A\right)}{\partial\alpha} < 0$, $\frac{\partial\left(q^P - q^A\right)}{\partial\beta} < 0$

The Agent's participation threshold $q^0(\widehat{\alpha}\beta)$ is the unique root of the equation

$$U(q;\widehat{\alpha},\beta) = B + \widehat{\alpha}\beta q - C(q) = 0 \tag{5}$$

This leads to the following Lemma.

Lemma 4. *The Agent's participation threshold is given by an increasing function* $q^0(\widehat{\alpha},\beta)$.

The proof of Lemma 4 is given in the Appendix.

3 Reciprocal Altruism and Contracts

In this section I build variants of the reciprocal altruism framework so as to account for the behavior observed in the two well-known lab experiments - the

Trust Game of Fehr and Rockenbach [2003] and the Control Game[7] by Falk and Kosfeld [2006], which are described in detail below.

Under the selfish players assumption, the results of the experiments are puzzling whereas the reciprocal altruism framework allows to account for the actually observed behavior, for instance, the motivation crowding out. This also shows the potential of the modeling approach based on reciprocal altruism and provides an additional justification for the relevance of contracting models based on it.

3.1 The Trust Game

Consider the Trust Game (or Investment Game) in its Fehr and Rockenbach [2003] version. In their experiment, both the Principal and the Agent are endowed with $S = 10$ units of money. First, the Principal decides on x - how much money to send to the Agent and also announces $\widehat{q}$ - the desired back-transfer. The desired back-transfer isn't binding for the Agent. The experimenter triples the sum of money sent by the Principal[8], so that the Agent receives $3x$. The Agent then decides on the back-transfer q. This setting represents the *Baseline treatment*. Notice that $\widehat{q}$ is a "cheap talk" in this case.

In the *Incentive treatment* the Principal on top of x and $\widehat{q}$ announces a fine f which is imposed on the Agent if the back-transfer is lower than the desired level $\widehat{q}$, so that $\widehat{q}$ is no more "cheap talk". The fine isn't paid to the Principal, it simply reduces the Agent's payoff, in other words the fine is a punishment for the Agent. The fine amount is exogenous (set by the experimenter), so that the only decision of the Principal is to choose whether to impose the fine or not. The contract is now $(x, \widehat{q}, f)$ where f can take only two values 0 or $\overline{f}$.

The paper brings evidence of crowding-out, i.e. the use of the extrinsic motivator (fine for bad performance) decreases the intrinsic motivation and, as a result, leads to a lower performance. The study finds that, on average, in the incentive treatment the back-payment is higher when the fine is set to zero (the Principal chooses not to punish) than for the case of punishing ($f = \overline{f}$). This means that imposing an extrinsic incentive leads to a lower performance.

The utilities of the Principal and the Agent in the experimental setting are given, following the reciprocal altruism framework[9] by

$$V = 10 - x + q + \alpha(10 + 3x - C(q) - fI_{q<\widehat{q}})$$
$$U = 10 + 3x - C(q) - fI_{q<\widehat{q}} + \widehat{\alpha}\beta\,(10 - x + q)$$

Suppose that the decision on x has already been made and focus on the continuation subgame[10] in which the Principal decides on $\widehat{q}$ and f and then the

[7] I follow Ellingsen and Johannesson [2008] in calling the game of Falk and Kosfeld [2006] the "Control Game".

[8] This explains why the game can also be called the "Investment Game". The transfer x can be thought of as an investment, $3x$ - as a return to the investment.

[9] In the experiment monetary cost of paying back is linear: $C_m(q) = q$. I assume that there is also a psychological cost of paying back $C_\psi(q)$ which assumed to be convex, so that the overall cost $C(q) = C_m(q) + C_\psi(q)$ is convex. This assumption is admittedly ad hoc, but it is needed to capture the predominance of non bang-bang behavior.
Alternatively, one can assume that Principal's utility from money is concave with linear cost. Then, after rescaling utility to linear, cost become convex.

[10] Of course, x itself is a signal of the Principal's altruism, but I assume that the Agent updates his belief on the Principal's altruism after observing x, which brings the belief at the beginning of the subgame.

Agent decides on q. We can consider x as a constant at this point. Dropping x and other constants in the payoff functions leads to the following simplified expressions for the payoffs:

$$V = q - \alpha(C(q) + fI_{q<\widehat{q}}) \tag{6}$$
$$U = \widehat{\alpha}\beta q - C(q) - fI_{q<\widehat{q}} \tag{7}$$

I specify now the distribution of the Principal' and the Agent's characteristics and information structure of the game.

Let the Principals and the Agents be heterogenous - some of them are pro-social, others are selfish. I denote the type of the Principal by θ^P, and the type of the Agent by θ^A. For both - the Principals and the Agents, $\theta^j \in \{\text{Social}, \text{Selfish}\}$. The type is the private information.

The pro-social actors are characterized by altruism α_H and reciprocity intensity β_H, the selfish ones - by the pair (α_L, β_L), where

$$\alpha_H > \alpha_L, \quad \beta_H > \beta_L, \quad 0 \leq \alpha_j \leq 1, \quad \alpha_H\beta_H \leq 1$$

To simplify the analysis, I assume[11] that $\beta_L = 0$.

Players (Principals and Agents) are drawn from the same population. The share of the pro-social actors is not known, but the actors have some beliefs about the population composition. Players rationally believe that the others in the society (or population) are like themselves, i.e. they exhibit rational projection bias[12] (i.e. consistent with Bayes rule). Loewenstein et al. [2003] provide evidence for the existence of projection bias and develop a formal model. Bénabou and Tirole [2006b] discuss the implication of the projection bias for collective beliefs.

Denote by π_H the probability, assigned by the pro-social Principal to being matched with the pro-social Agent and by π_L the probability, assigned to the same event by the selfish Principal:

$$\pi_H = Prob(\theta^A = \text{Social}|\theta^P = \text{Social}) = Prob(\beta = \beta_H|\alpha = \alpha_H) \tag{8}$$
$$\pi_L = Prob(\theta^A = \text{Social}|\theta^P = \text{Selfish}) = Prob(\beta = \beta_H|\alpha = \alpha_L) \tag{9}$$

According to the rational projection bias, assume that

$$\pi_L < \pi < \pi_H$$

where π is the true share of the pro-social actors.

Notice that the Principal moves first and doesn't know the type of the Agent with whom she is matched. The Agent, on the contrary, observes the action of the Principal, and can use this to learn about the Principal's type. Because of this, I suppose that behavior of the Principal is driven by her (unconditional) altruism[13], whereas the behavior of the Agent is driven by his reciprocity, which

[11] The more general setting with the four possible pairs (α_k, β_l) and without requiring $\beta_L = 0$ can be considered. This, however, leads to more tedious computations but doesn't bring additional intuition. So, I restrict attention to the simpler setting.

[12] The rational projection bias can be determined as "tendency to look at others (other people or future selves) from the point of view of one's current self" - see Tirole [2002].

[13] Alternatively, one can assume that given the prior belief on the Agent's altruism, the Principal's altruism is equal to the sum of her pure (unconditional) altruism α_p and reciprocal altruism $\alpha_r = \beta_P E[\alpha^A]$. This results in the Principal's altruism towards the Agent at the level $\alpha_H = \alpha_{pH} + \beta_H E[\alpha^A]$ or $\alpha_L = \alpha_{pL} + \beta_L E[\alpha^A]$, depending on the type of the Principal.

is reflected by the structure of altruism in the utility functions V and U in (6) and (7): only altruism α affects the Principal's behavior, and only reciprocity intensity β, interacted with the belief on the Principal's type $\widehat{\alpha}$ - affects the Agent.

This setting brings us to the following signaling game with two-sided asymmetric information.

The Principal is of type $\theta^P \in \{\text{Social, Selfish}\}$, or, equivalently, $\alpha \in \{\alpha_H, \alpha_L\}$. The Agent is of type $\theta^A \in \{\text{Social, Selfish}\}$, or, equivalently, $\beta \in \{\beta_H, \beta_L\}$. Each player's type is his/her private information.

To make notation simpler, I will write in what follows that the Principal is of type $i = H$ ($i = L$) when $\theta^P = Social$ ($\theta^P = Selfish$), which is equivalent to $\alpha = \alpha_H$ ($\alpha = \alpha_L$). Similar convention will be used for the type j of the Agent.

The Principal's strategy is a type-contingent pair $(f_i, \widehat{q}_i) \in \{0, \overline{f}\} \times [0, +\infty)$, $i = L, H$ (where the index L is used for the selfish type, and H indexes for the pro-social type).

The Agent's strategy is a type-contingent back-transfer conditional on the Principal's action $q_j(f, \widehat{q})$ where $q_j \in [0, +\infty)$, $j = L, H$.

The Principal has belief on the probabilities to be matched with the pro-social Agent, dependent on the Principal's type: π_H or π_L.

The Agent's ex-post belief μ is determined by the Principal's observed action, $\mu(f, \widehat{q}) = Prob(i = H|f, \widehat{q})$. There is a one-to-one correspondence between belief μ and the ex-post expectation of the Principal's type $\widehat{\alpha}$: $\widehat{\alpha} = \mu\alpha_H + (1-\mu)\alpha_L$, so that $\widehat{\alpha}$ can be considered instead of μ. The payoffs are given by (6) and (7).

The solution concept is the Perfect Bayesian equilibrium, in which Agent's belief off the equilibrium path are "reasonable", in the sense of the intuitive criterion of Cho and Kreps.

The game described above corresponds to the Incentive Treatment. For the Baseline Treatment the fine f is exogenously set to zero.

I now proceed backwards in the analysis of the game.

Consider the Agent's Best Response back-transfer for the different treatments of the experiment. The Agent's participation threshold isn't relevant, since paying back zero is feasible.

Claim 1. *In the Trust Game, if the Agent holds beliefs $\widehat{\alpha}$ on the Principal's altruism, the Best Response back-transfer q is:*

1. *in the baseline treatment (fine isn't possible) and in the incentive treatment when the Principal chooses not to punish ($f = 0$): $q = q^A(\widehat{\alpha}\beta)$.*

2. *in the incentive treatment when the Principal chooses to impose a fine ($f = \overline{f}$):*

$$q = \begin{cases} q^A(\widehat{\alpha}\beta) & \text{if } \widehat{q} < q^A(\widehat{\alpha}\beta) \\ \widehat{q} & \text{if } q^A(\widehat{\alpha}, \beta) < \widehat{q} < \widetilde{q}^A(\widehat{\alpha}, \beta) \\ q^A(\widehat{\alpha}\beta) & \text{if } \widehat{q} > \widetilde{q}^A(\widehat{\alpha}\beta) \end{cases}$$

where $\widetilde{q}^A(\widehat{\alpha}\beta)$ is determined by (see figure 2)

$$\widehat{\alpha}\beta q^A - C(q^A) - f = \widehat{\alpha}\beta\widetilde{q}^A - C(\widetilde{q}^A), \quad \text{and} \quad \widetilde{q}^A > q^A$$

and $\widetilde{q}^A(\widehat{\alpha}, \beta)$ is an increasing function of $\widehat{\alpha}$.

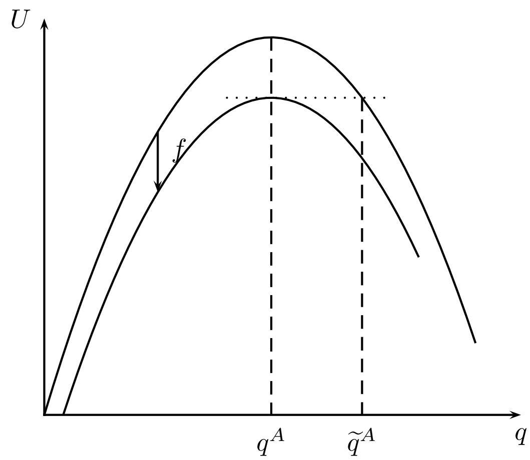

Figure 2: Agent's payoff

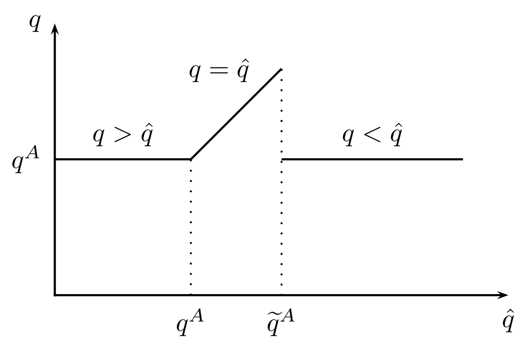

Figure 3: Back-transfer as a function of threshold

The proof of Claim 1 is given in the Appendix.

The value $\widetilde{q}^A(\widehat{\alpha}\beta)$ can be interpreted as the maximal performance level, which can be implemented when extrinsic incentives are provided to the intrinsically motivated Agent, whereas $q^A(\widehat{\alpha}\beta)$ is the maximal value which can be implemented with intrinsic motivation only. For a given belief $\widehat{\alpha}$ holds $\tilde{q}^A(\widehat{\alpha}\beta) > q^A(\widehat{\alpha}\beta)$, and so under symmetric (or revealed) information, if an extrinsic incentive is added to intrinsic motivation, the performance level is higher. So, the fine serves as a "positive reinforcer".

When the threat of the fine is imposed, the Agent has to give up some utility and faces a trade-off between complying and departing from his preferred back-payment q^A by selecting the desired back-payment $\widehat{q}$ and paying a fine and sticking with his preferred back-payment q^A. The value $\widetilde{q}^A$ determines the Agent's preferred option: if the desired back-payment is below this value, the Agent prefers to diverge from q^A to $\widehat{q}$, if the desired back-transfer is higher than $\widetilde{q}^A$, the Agent prefers to disobey and pay the fine. So, $q > \widetilde{q}^A$ can't be implemented with a threat of fine.

It follows from the Claim that, contrary to the standard contract theory, when extrinsic incentives are used for the intrinsically motivated Agent, the actual back transfer can be higher, equal or lower than the desirable back-transfer, as illustrated by Figure 3.

If $\widehat{q} < q^A$, the required performance is low compared to the intrinsic motivation, so the Agent is willing to perform better than he is asked for. For $\widehat{q} = q^A$ the intrinsic motivation is just enough to motivate the Agent for the required level of performance. Finally, $\widehat{q} > q^A$ corresponds to the case when intrinsic motivation isn't enough to inspire the Agent for high enough performance.

Before starting the analysis of the Principal's move, I introduce some notation.

Denote by

$$q_{ij} = q^A(\alpha_i, \beta_j)$$

the maximal back-transfer which can be implemented by using intrinsic motivation only, given that the Agent with $\beta = \beta_j$ holds belief $\widehat{\alpha} = \alpha_i$. These back-transfers are determined by $C'(q_{ij}) = \alpha_i \beta_j$. Since $\beta_L = 0$,

$$q_{HL} = q_{LL} = 0$$

Denote by

$$\widetilde{q}_{ij} = \widetilde{q}^A(\alpha_i, \beta_j)$$

the maximal back-transfer which can be implemented by using both intrinsic and extrinsic motivation, i.e. by imposing the (threat of) fine, given that the Agent with $\beta = \beta_j$ holds belief $\widehat{\alpha} = \alpha_i$. It follows from Claim 1 that

$$C(\widetilde{q}_{LL}) = f \tag{10}$$

I now show that under some restrictions on the parameters, the game has the separating equilibrium with crowding-out.

Proposition 1. *Assume that* $q_{LH} \leq \widetilde{q}_{LL}$.

If

$$\widetilde{q}_{LH} \leq q_{HH}, \qquad \pi_H \geq \widehat{\pi}_H, \qquad \pi_L \leq \widehat{\pi}_L \tag{11}$$

where

$$\widehat{\pi}_H = \frac{\widetilde{q}_{LL} - \alpha_H C(\widetilde{q}_{LL})}{q_{HH} - \alpha_H C(q_{HH})} \leq 1 \tag{12}$$

$$\widehat{\pi}_L = \frac{\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})}{q_{HH} - \alpha_L C(q_{HH})} \geq 0 \tag{13}$$

then there exists unique equilibrium satisfying the Cho-Kreps intuitive criterion. The equilibrium is separating and has the crowding-out property.

The conditions (11) *are also necessary for the existence of the crowding-out equilibrium.*

In this equilibrium the α_H*-type imposes no (threat of a) fine, the* α_L*-type threatens a fine:*

$$f^*_L = \overline{f}, \quad f^*_H = 0$$

The desired back-transfers are: $\widehat{q}^*_L = \widetilde{q}_{LL}$, *any* $\widehat{q}^*_H \leq q_{HH}$.

In particular, it is possible to have

$$\widehat{q}^*_L > \widehat{q}^*_H$$

The actual back-transfers are

$$q^*_{LL} = \widetilde{q}_{LL}, \quad q^*_{LH} = \widetilde{q}_{LL}, \quad q^*_{HL} = 0, \quad q^*_{HH} = q_{HH}$$

where q^*_{ij} is the equilibrium back-transfer to the α_i-Principal from the β_j-Agent.

The average back-transfer to the pro-social Principal is higher than that to the selfish one:

$$\pi q^*_{HH} \geq q^*_{LL}$$

The Proof of Proposition 1 is given in the Appendix.

The assumption $q_{LH} \leq \widetilde{q}_{LL}$ means that when the selfish Principal controls, revealing his type in this way, both types of Agents exert the minimal effort $\widetilde{q}_{LL}$. Were the inequality $q_{LH} \leq \widetilde{q}_{LL}$ to be reversed, the pro-social Agent would exert effort q_{LH} when controlled. In this case the proposition still holds with the thresholds $\widehat{\pi}_j$ determined by other (more complicated) formulae.

The condition $\widetilde{q}_{LH} \leq q_{HH}$ is crucial to guarantee crowding-out. In fact, were it be violated, the separating equilibrium could still emerge, in which case[14] the selfish Principal controls at level $\widetilde{q}_{LH}$ and obtains this effort from the pro-social Agent. Since the pro-social Principal obtains the lower effort q_{HH} from the pro-social Agent, the average performance for her is lower compared to the selfish Principal (the selfish Agents exert zero effort in both cases - whether they are controlled or not). Finally, the conditions on beliefs π_j are equivalent to the incentive compatibility constraints for a separating equilibrium.

To guarantee that (12) and (13) in Proposition 1 hold, the sorting condition is required. For a given setting it can be written as

$$\pi_H V_H(q_T) - V_H(q_P) \geq \pi_L V_L(q_T) - V_L(q_P) \tag{14}$$

for all $q_P \leq q_T \leq Q$, where $V_i(q) = V(q, \alpha_i)$, q_P stands for performance, realized when the Principal chooses to "punish", i.e. $f = \overline{f}$, q_T stands for performance, realized when the Principal chooses to "trust", i.e. $f = 0$. The cut-off $Q = \max\{q_{HH}, \widetilde{q}_{LH}\}$ is needed because the utility function V is non-monotone.

The sorting condition says that when the pro-social Principal prefers to punish, the selfish one prefers to punish even stronger. Whereas, if the selfish Principal prefers to trust, the pro-social one prefers to trust to a larger degree.

The proposition makes it clear that the emerging separating equilibrium in the signaling game accounts for the behavioral patterns observed in the experiment. Intuitively, the pro-social Principal chooses not to impose the fine, i.e. not to use the extrinsic incentive, signaling in this way her generosity and inspiring high intrinsic motivation of the pro-social Agent, which perform at a comparatively high level q_{HH}. If the pro-social Principal is matched with a selfish Agent, who is not intrinsically motivated, the performance is zero: $q_{LL} = 0$, since there are no extrinsic incentives.

Imposing the threat of fine (negative "bonus for bad performance"), i.e. using the extrinsic incentive, crowds out the intrinsic motivation of the pro-social Agent because this signals low altruism of the Principal. However, provision of the extrinsic incentive guarantees performance of all Agents[15] at a comparatively low level $\widetilde{q}_{LL}$. This can lead to the observed crowding out in the Agent's

[14]For some values of parameters there can be pooling equilibria.

[15]The pro-social Agents could perform at the level q_{LH}, if $q_{LH} > \widetilde{q_{LL}}$. This, however, ruled out by the Proposition assumption.

performance: the average for the lottery between q_{HH} and $q_{LL} = 0$ is higher than $\widetilde{q}_{LL}$ for some values of the parameters.

The rational projection bias is crucial to guarantee the no-mimicking (incentive compatibility) condition. The Principals have different beliefs on the Agent's type, for instance, the pro-social Principal is more optimistic. Because of this, the pro-social Principal prefers the lottery between high q_{HH} and zero, whereas the selfish Principal prefers the comparatively low but sure outcome[16] $\widetilde{q}_{LL}$.

I now turn to a more structured discussion of the conditions for the separating crowding out equilibrium to emerge. The condition $q_{LH} \leq \widetilde{q}_{LL}$ is equivalent to $C(q_{LH}) \leq C(\widetilde{q}_{LL})$. Since $C(\widetilde{q}_{LL}) = \overline{f}$ - see (10), this means that $\overline{f}$ should be high enough:

$$\overline{f} \geq f_1 \equiv C(q_{LH})$$

The back-transfer $\widetilde{q}_{LH}$ is determined, according to Claim 1 by

$$\alpha_L \beta_H q_{LH} - C(q_{LH}) - \overline{f} = \alpha_L \beta_H \widetilde{q}_{LH} - C(\widetilde{q}_{LH})$$

where $\widetilde{q}_{LH}$ is chosen on the decreasing part of the graph of the function $F(q) = \alpha_L \beta_H q - C(q)$ (see Figure 2). Consequently, $\widetilde{q}_{LH} \leq q_{HH}$ is equivalent to

$$\alpha_L \beta_H q_{LH} - C(q_{LH}) - \overline{f} \geq \alpha_L \beta_H q_{HH} - C(q_{HH})$$

which can be rewritten as

$$\overline{f} \leq f_2 \equiv (\alpha_L \beta_H q_{LH} - C(q_{LH})) - (\alpha_L \beta_H q_{HH} - C(q_{HH}))$$

The conditions $\pi_H \geq \widehat{\pi}_H$ and $\pi_L \leq \widehat{\pi}_L$ with the thresholds given by (12) and (13), show that the projection bias should be large enough to ensure incentive compatibility.

The following corollary gives the more structured description of the crowding-out separating equilibrium.

Corollary 1. *For a generic triple $(\alpha_L, \alpha_H, \beta_H)$ and quadratic cost function $C(q) = \frac{c}{2}q^2$ there always exists a non-empty set of the parameters $(\pi_L, \pi_H, \overline{f})$ such that there exists a separating equilibrium with crowding-out.*

These parameters satisfy to (12) and (13) and the inequality $f_1 \leq \overline{f} \leq f_2$. If cost isn't quadratic, the additional condition for a triple $(\alpha_L, \alpha_H, \beta_H)$

$$\alpha_L \alpha_H \beta_H^2 \left(\frac{q_{LH}}{\alpha_L \beta_H} - \frac{q_{HH}}{\alpha_H \beta_H} \right) \leq C(q_{LH})$$

is required to obtain a non-empty of the parameters $(\pi_L, \pi_H, \overline{f})$, leading to a separating equilibrium with crowding-out.

[16]This reasoning can also be applied if the Principals and the Agents are drawn from two independent distributions, so that there is no projection bias, but the Principals are risk-averse. In this case, for the lottery between $V(q_{HH})$ and $V(0)$ the α_L-type has a higher spread, since $V_L(q_{HH}) > V_H(q_{HH})$ and $V_L(0) < V_H(0)$ (in the separating equilibrium $V_L(0) = -\alpha_L f$, $V_H(0) = 0$). Because of this, for some range of parameters, the certain outcome $V_L(\widetilde{q}_{LL})$ is better than the certainty equivalent of the lottery for the α_L-type, but below the certainty equivalent for the α_H-type. So, the separating equilibrium with crowding-out can also emerge in such setting.

The Proof of Corollary 1 is given in the Appendix.

The Corollary describes a set of parameters, under which the sorting condition (14) holds. It stays, for instance, that the available extrinsic incentive f should not be neither too weak neither too strong for the separating equilibrium with crowding-out to appear. Intuitively, if the available extrinsic motivator is too weak and its use reveals the selfish type, the Agent's intrinsic motivation is crowded-out, whereas extrinsic incentives is weak, which leads to a low performance. Because of this, such outcome can't be an equilibrium, as selfish Principal prefers not to reveal her type. On the other hand, if the available extrinsic motivator is too strong, the pro-social Principal prefers to use it, since it overcomes the decrease in intrinsic motivation, and pooling with threat of punishment becomes an equilibrium outcome.

3.2 The Control Game

In the experiment conducted in Falk and Kosfeld [2006] the Principal chooses whether to restrict the set of Agent's effort from below. Output is again equal to effort.

Put formally, the Principal offers a contract $\underline{q}$ which can take two values - 0 or q_c, where the latter is exogenously set by the experimenter. The Agent then chooses effort $q \in [\underline{q}, \infty)$. Effort is costly for the Agent. The Agent has an initial endowment of 120.

The experiment has a number of findings which can not be explained within the selfishness framework. For instance, the Agents, when offered a contract $\underline{q} > 0$, exert, on average, less effort, than when offered $\underline{q} = 0$, which means that extrinsic incentive (control) has a negative impact on Agents' performance. This demonstrates a hidden cost of control effect. The observed behavior of the Agents is heterogenous: there is observed positive, negative and neutral reaction to control. Finally, many Principals choose not to control.

The reciprocal altruism framework accounts for these experimental findings.

As for the Trust Game, I use the reciprocal altruism framework and build a model matching the experiment design.

The selfish utilities of the Principal and the Agent are given by[17]

$$v = q$$
$$u = 120 - C(q)$$

The reciprocal altruism framework leads to the (social) utilities

$$V = q + \alpha(120 - C(q))$$
$$U = 120 - C(q) + \widehat{\alpha}\beta q$$

The initial endowment of the Agent allows to disregard the Agent's participation constraint. By dropping the constants, the Principal's and Agent's utilities can be simplified to

$$V = q - \alpha C(q) \tag{15}$$
$$U = \widehat{\alpha}\beta q - C(q) \tag{16}$$

[17] The experiment sets $C(q) = q/2$. As for the Trust Game, I assume that $C(q)$ is convex. See footnote 9 for the justification of the assumption.

Denote by $q^A(\widehat{\alpha}\beta)$ the Agent's preferred effort level as in the benchmark case - see (4).

Consider the setting with heterogenous Principals and Agents, adopted in the analysis of the Trust Game in subsection 3.1.

The Principal's strategy is a type-contingent choice of control $\underline{q}_i \in \{0, q_c\}$, $i = L, H$. The Agent's strategy is a type-contingent effort, conditional on the Principal's action $q_j(\underline{q}) \in [\underline{q}, +\infty)$, $j = L, H$.

The Principal's belief on the probabilities to be matched with the pro-social Agent depends on the Principal's type and are given by (8) and (9). The Agent's ex-post beliefs are determined by the Principal's observed action, $\mu(\underline{q}) = Prob(\alpha = \alpha_H | \underline{q})$. There is a one-to-one correspondence between belief μ and the ex-post expectation of the Principal's type $\widehat{\alpha}$: $\widehat{\alpha} = \mu\alpha_H + (1-\mu)\alpha_L$, so that $\widehat{\alpha}$ can be considered instead of μ. The payoffs are given by (15) and (16).

As in the analysis of the Trust game, I look for a Perfect Bayesian equilibrium in which Agent's beliefs off the equilibrium path are "reasonable" in the sense of the intuitive criterion of Cho and Kreps.

I proceed backwards in the analysis of the game.

Consider first the Agent's Best Response choice of effort.

Claim 2. *If $q^A(\widehat{\alpha}, \beta) \geq \underline{q}$ then the Agent's Best Response is $q = q^A(\widehat{\alpha}, \beta)$; otherwise it is $q = \underline{q}$.*

The Claim is evident as it simply says that the Agent chooses the global maximizer of his utility whenever it's feasible. Otherwise, he chooses the closest feasible effort which is equal to $\underline{q}$ - the lower bound of the set of feasible efforts.

Denote by q_{ij} the effort, voluntarily exerted by the Agent with β_j which beliefs that the Principal's type is α_i, i.e. $q_{ij} = q^A(\alpha_i, \beta_j)$ and $C'(q_{ij}) = \alpha_i\beta_j$.

Consider now the Principal's decision. If the Principal with altruism α_i holds beliefs π_i to be matched with a pro-social Agent, who, in turn, holds the true beliefs about the Principal's type, then under no-control, the Principal's utility is[18]

$$V = \pi_i(q_{iH} - \alpha_i C(q_{iH})) \tag{17}$$

Under control, if effort q_{iH} is available, i.e. $q_{iH} \geq q_c$, and the Principal's utility is still given by (17). If $q_{iH} < q_c$, then the effort $q = q_c$ is implemented and the Principal's utility is

$$V = (q_c - \alpha_i C(q_c)) \tag{18}$$

Denote by $q_i^{C1} < q_i^{C2}$ the two roots of the equation (see Figure 4)

$$\pi_i(q_{iH} - \alpha_i C(q_{iH})) = (q_c - \alpha_i C(q_c))$$

Comparing the Principal's utility for the case when the Agent holds the true beliefs on the Principal's type, given by (17) and (18), leads to the following characterization of the Principal's Best Response in this case:

Claim 3. *Under symmetric information about the Principal's type, the type α_i -Principal's optimal strategy is:*

$$\underline{q} = \begin{cases} q_c & \text{if } q_c \leq q_i^{C2} \\ 0 & \text{if } q_c \geq q_i^{C2} \end{cases}$$

[18] $q_{iL} = 0$ since $\beta_L = 0$.

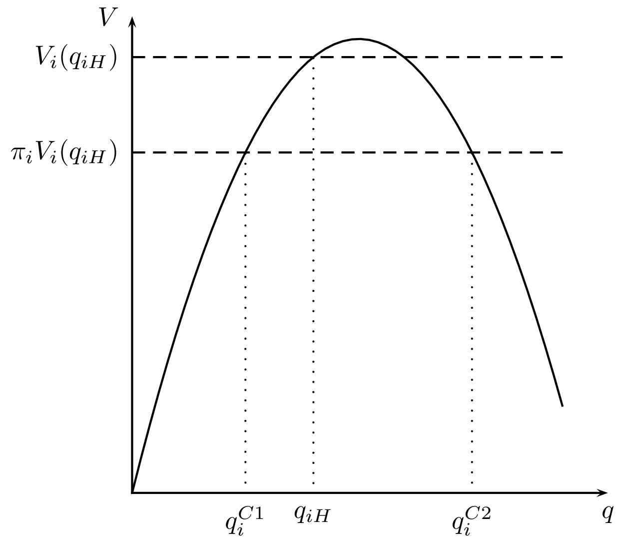

Figure 4: Principal's utility

Consider now the case when the Principal's type is her private information.

Proposition 2. *For each $q_c < q_{HH}$, $(\alpha_L, \alpha_H, \beta_H)$ there exists a range of parameters (π_L, π_H) such that there exists a unique separating equilibrium of the Control Game satisfying the intuitive criterion. There is crowding-out in effort in equilibrium.*

The parameters should be such that

$$\pi_L \leq \widehat{\pi}_L, \quad \pi_H \geq \widehat{\pi}_H$$

where for the case $q_c \geq q_{LH}$

$$\widehat{\pi}_L = \frac{q_c - \alpha_L C(q_c)}{q_{HH} - \alpha_L C(q_{HH})} > 0$$
$$\widehat{\pi}_H = \frac{q_c - \alpha_H C(q_c)}{q_{HH} - \alpha_H C(q_{HH})} < 1$$

for the case $q_c < q_{LH}$

$$\widehat{\pi}_L = \frac{q_c - \alpha_L C(q_c)}{[q_{HH} - \alpha_L C(q_{HH})] + [q_c - \alpha_L C(q_c)] - [q_{LH} - \alpha_L C(q_{LH})]} > 0$$
$$\widehat{\pi}_H = \frac{q_c - \alpha_H C(q_c)}{[q_{HH} - \alpha_H C(q_{HH})] + [q_c - \alpha_H C(q_c)] - [q_{LH} - \alpha_H C(q_{LH})]} < 1$$

In equilibrium, the pro-social Principal doesn't control, whereas the selfish Principal does:

$$\underline{q}_H^* = 0, \quad \underline{q}_L^* = q_c$$

The Agent's performance

$$q_{HH}^* = q_{HH}, \quad q_{HL}^* = 0, \quad q_{LH}^* = \max\{q_{LH}, q_c\}, \quad q_{LL}^* = q_c$$

where q^*_{ij} is the equilibrium back-transfer to the α_i-Principal from the β_j-Agent.

The average effort to the pro-social Principal is higher than that to the selfish one.

The Proof of Proposition 2 is given in the Appendix.

As for the Trust Game, the sorting condition, similar to (14) is required. It is written as

$$\pi_H V_H(q_T) - V_H(q_c) > \pi_L V_L(q_T) - V_L(q_c) \tag{19}$$

for the case of $q_c \geq q_{LH}$, and

$$\pi_H V_H(q_T) - [\pi_H V_H(q_c) + (1-\pi_H)V_H(q_{LH})] > \pi_L V_L(q_T) - [\pi_L V_L(q_c) + (1-\pi_L)V_H(q_{LH})] \tag{20}$$

for the case of $q_c < q_{LH}$.

Since the Principal has only binary choice - to control or to trust, the sorting condition should be verified only for $q_T = q_{HH}$ (an outcome q_c or a lottery between q_c and q_{LH} correspond to a (variable) outcome q_P in (14)).

The mechanism of emerging of the separating equilibrium with crowding-out is similar to that of the Trust Game. By choosing not to control, the pro-social Principal signals her kindness, and this inspires high intrinsic motivation for the pro-social Agent. Because of this, when matched with the pro-social Principal, the pro-social Agent exerts high effort q_{HH}. However, the selfish Agent doesn't react to the signal of the Principal's generosity and, once not controlled, exerts zero effort.

The selfish Principal chooses to control and guarantees the (comparatively low) output q_c. However, even selfish Principal inspires the pro-social Agent's intrinsic motivation, so effort from the pro-social Agent can be q_{LH}, if it's higher than the controlling threshold q_c.

The lottery between q_{HH} and 0 effort is differently treated by the two Principals. The pro-social one is more optimistic, and beliefs that the chance to get q_{HH} is higher, compared to the beliefs of the selfish Principal. Because of this, no-mimicking holds - the pro-social Principal prefers the lottery, whereas the selfish Principal prefers the sure outcome.

As in the Trust Game, the separating crowding-out equilibrium emerges when the available extrinsic incentive is neither too weak nor too strong.

I now turn to the more detailed description of other equilibrium structures of the Control Game.

Proposition 3. *For given $\alpha_L, \alpha_H, \beta_H, \pi_L, \pi_H$, there exist the threshold values q_i, $q_i < q_j$ for $i < j$, such that the equilibrium in the Control Game is:*

1. *No-control pooling for $q_c \in [0, q_1]$, which represents no effect of control;*
2. *Separating equilibrium with crowding-out for $q_c \in [q_2, q_3]$ (hidden cost of control);*
3. *Control pooling for $q_c \in [q_3, q_4]$ (positive effect of control);*
4. *Separating with no crowding-out in effort $q_c \in [q_4, q_5]$ (positive effect of control);*
5. *No-control pooling $q_c \in [q_6, +\infty)$.*

For $q_c \in [q_1, q_2]$ and $q_c \in [q_5, q_6]$ an equilibrium involves mixed strategies.

The Proof of Proposition 3 is given in the Appendix.
Figure 5 illustrates the proposition.

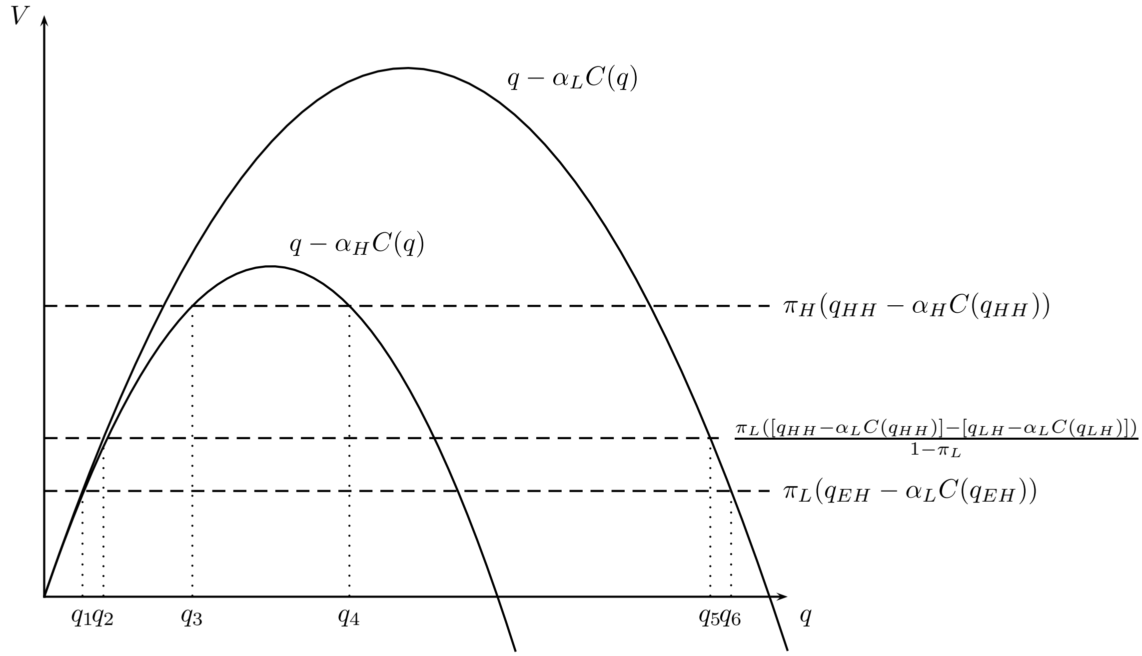

Figure 5: Equilibrium structure in the Control Game

Falk and Kosfeld [2006] found in the experiment that for small q_c the hidden cost of control effect is stronger compared to large q_c. The model is in line with this result. In fact, the hidden cost of control is obtained for $q_c \in [q_2, q_3]$; for larger q_c, the pooling equilibrium with control emerges which means that the increase in q_c will lead to the increase in average performance (which is equal to q_c). For even larger q_c the separating equilibrium in which the controlling Principal gets larger output than the non-controlling one, so there is also the positive effect of control rather than the hidden cost.

The experiment finds strong heterogeneity between the Agents and that there are Agents which react positively, neutrally or negatively to control. The heterogeneity of the Agents is assumed in the model, and the reaction of the Agents is predicted. Controlling the pro-social Agents can lead to the decrease in his performance. For instance, for $q_c \in [q_2, q_3]$ when the crowding-out equilibrium emerges, the controlled pro-social Agent performs at $q = q_c$, whereas the uncontrolled at $q = q_{HH} > q_c$. Controlling the selfish Agent leads to the increase in performance for $q_c \in [q_2, q_5]$. Controlling the pro-social Agents also leads to the increase in their performance for the control pooling equilibria. Finally, the neutral effect of control can emerge in the game but out of equilibrium path. According to Claim 2, if Agent's preferred output q^A is higher than control, the Agent will perform at level q^A independently on whether he is controlled or not. In particular, for the case of weak control (q_c is small enough), even if the pro-social Agent is controlled, he can choose to perform at a higher level.

The experiment finds that for larger q_c, the larger share of the Principals chooses to control. This is also the prediction of the model: for small q_c none of the Principals controls which results in the no-control pooling equilibrium; for larger values of q_c the hidden cost of control equilibrium emerges in which only the selfish Principals control; for even larger values of q_c the control pooling equilibrium emerges in which both types of Principal choose to control.

4 Contracts without Extrinsic Incentives

In this section I consider the case when extrinsic incentives are not available at all. In this case the Agent is motivated to exert effort either through his intrinsic motivation either because he's "forced" to perform by the contract. In the former case the Agent is happy (gets positive utility), in the latter case the Agent doesn't enjoy the interaction with the Principal (i.e. performs at his participation constraint).

Let the Principal's altruism parameter α be her private information and the rest be symmetrically known.

The timing is as follows:

1. Principal learns α.
2. Principal offers a contract[19]: a command, i.e. specifies the output q.
3. Agent accepts or rejects the contract.
4. Contract is implemented and payoffs are realized.

So, we have a signaling game with common values.

To simplify the exposition even further, I consider the two-type case and then generalize the result to the continuum of types case.

Even though the formal setting of the signaling game should be clear for most of the readers, I provide its formal description in the next subsection and proceed then to the analysis of Perfect Bayesian Equilibria and refinement.

Readers who are not interested in the technical details can skip the technical subsection 4.1 and jump to 4.2 where the outcome of the signaling game as predicted by the refined equilibrium is described.

4.1 Signaling with 2 types

There are 2 players - Principal (sender) and Agent (receiver).

The Principal's type is her private information. Denote by $\mathcal{A}$ the set of the possible types, $\mathcal{A} = \{\alpha_H, \alpha_L\}$. The prior distribution is given by

$$\alpha = \begin{cases} \alpha_H & \text{with prob. } \Pi \\ \alpha_L & \text{with prob. } 1 - \Pi \end{cases} \tag{21}$$

The set of actions for the Principal[20] is $Q = [0, +\infty)$. The set of actions for the Agent is $A = [0, 1]$ with $a \in A$ be the probability of acceptance of an offer made by the Principal.

[19] The contract is the take-it-or-leave-it offer.

[20] Actually, this set of actions can be reduced to $[0, q_H^0]$.

A pure strategy of the Principal is a type-contingent $q \in Q$, i.e. q_H for α_H type and q_L for α_L-type.

A pure strategy of the Agent is an acceptance rule $a(\cdot)$. The value $a(q) \in [0, 1]$ is the probability of accepting the offer q. The set of Agent's pure strategies $\mathcal{F}$ is the set of all mappings from Q to $[0, 1]$.

A mixed strategy of the Principal is a probability distribution over Q conditional on type, $\sigma(\cdot|\alpha)$. Clearly her mixed strategy can be represented by the two probability distributions over Q: $(\sigma_H(\cdot),\sigma_L(\cdot))$.

A mixed strategy of the Agent is a probability distribution[21] τ over $\mathcal{F}$. The resulting mapping is still a mapping from Q to $[0, 1]$. So, the mixed strategy of the Agent, denoted by $a_\tau(\cdot)$, is still an element of $\mathcal{F}$. I will will restrict attention to the pure strategies of the Agent.

The Agent's ex-post beliefs on the Principal's type distribution $\mu(\cdot)$ is

$$\mu(q) = Prob(\alpha = \alpha_H|q)$$

There is one-to-one correspondence between Agent's beliefs $\mu(q)$ and Agent's ex-post expectation on the Principal's altruism

$$\widehat{\alpha}(q) = \mu(q)\alpha_H + (1 - \mu(q))\alpha_L \tag{22}$$

which is paralleled in the Agent's ex-post expected payoff[22]

$$U_\mu(q, a; \beta) = U(q, a, \widehat{\alpha}; \beta)$$

The pure-strategy profile is thus $((q_H, q_L), a(\cdot), \mu(\cdot))$. The mixed strategies-beliefs profile is $((\sigma_H(\cdot), \sigma_L(\cdot)), a(\cdot), \mu(\cdot))$.

The payoffs in the game for the pure-strategy profile $((q_L, q_H), a(\cdot))$ are

$$V(q, a(\cdot), \alpha) = V(q, \alpha)a(q) = (q - \alpha C(q) + \alpha B)a(q)$$
$$U(q, a(\cdot), \alpha; \beta) = U(q, \alpha; \beta)a(q) = (B - C(q) + \alpha\beta q)a(q)$$

which parallel (1) and (2).

The payoffs for the mixed strategies can be determined in a standard way.

Denote by q_H^P and q_L^P the preferred output of the high and low altruism Principals, respectively. Formally, $q_H^P = q^P(\alpha_H)$, $q_L^P = q^P(\alpha_L)$. From lemma 1 we have

$$q_H^P < q_L^P \tag{23}$$

Intuitively, the Principal who cares more about the Agent wants him to work less. Intuitively, since the marginal cost of effort is increasing and the Principal partially internalizes this cost, the one with stronger internalization prefers to have a lower marginal cost $C'(q)$ because marginal benefit from output is constant (equal to 1).

Denote by q_H^0 and q_L^0 the Agent participation thresholds when he learns that the Principal type is α_H and α_L correspondingly. Formally, $q_H^0 = q^0(\alpha_H, \beta)$, $q_L^0 = q^0(\alpha_L, \beta)$.

Denote the participation threshold when there is no update on the Principal's type by q_E^0:

$$q_E^0 = q^0(E\alpha, \beta)$$

[21] There is an issue of measurability over the space of functions.

[22] It is easy to see that $U_\mu(q, a(\cdot); \beta) = \mu(q)(B - C(q) + \alpha_H\beta q))a(q) + (1 - \mu(q))(B - C(q) + \alpha_L\beta q))a(q) = (B + \widehat{\alpha}\beta q - C(q))a(q)$.

where

$$E\alpha = \Pi\alpha_H + (1-\Pi)\alpha_L$$

According to Lemma 4,

$$q_L^0 < q_E^0 < q_H^0. \tag{24}$$

Intuitively, if the Principal's type is revealed to the Agent, then he is willing to exert more effort for a Principal who cares more about him. This is natural since the worker internalizes the benefits from output and intensity of the internalization is higher for the worker connected with more altruistic Principal.

4.1.1 The Perfect Bayesian Equilibrium

As is standard in signaling games, the set of the Perfect Bayesian Equilibria (PBE) is large. In this part of the paper I characterize the structure of the PBE set and proceed to refinement in section 4.1.2.

Consider a PBE of the signaling game. Denote the equilibrium (pure) offers of the Principal of type α_H (α_L) by q_H^* (q_L^*), the equilibrium acceptance rule of the Agent by $a^*(\cdot)$. The beliefs supporting the equilibrium is $\mu^*(\cdot)$. So, the pure-strategy equilibrium profile is $((q_H^*, q_L^*), a^*(\cdot); \mu^*(\cdot))$. Similarly, the mixed-strategy equilibrium profile is $((\sigma_H^*(\cdot), \sigma_L^*(\cdot)), a^*(\cdot); \mu^*(\cdot))$.

Denote by $BR_\mu(q)$ the best response acceptance rule for the Agent with ex-post beliefs $\mu(\cdot)$:

$$BR_\mu(q) = \arg\max_{a\in[0,1]} U_\mu(q, a; \beta) \tag{25}$$

Lemma 5. *For any beliefs $\mu(\cdot)$, the Best Response acceptance rule is a threshold with the threshold value*

$$\widehat{q}(q) = q^0(\widehat{\alpha}(q), \beta) \tag{26}$$

$$BR_\mu(q) = \begin{cases} 1 & \text{if } q < \widehat{q}(q) \\ 0 & \text{if } q > \widehat{q}(q) \\ \text{any } a \in [0,1] & \text{if } q = \widehat{q}(q) \end{cases}$$

where $\widehat{\alpha}(q)$ is given by (22)

For any beliefs $\mu(\cdot)$ and any offer q

$$q_L^0 \le \widehat{q}(q) \le q_H^0 \tag{27}$$

so that for any Best Response acceptance rule, the Agent accepts at least offers $q < q_L^0$ and rejects any offer $q > q_H^0$.

Corollary 2. *The* equilibrium *acceptance rule $a^*(\cdot)$ is a threshold with threshold value given by (26)*

The Proof of Corollary 2 is given in the Appendix.

I now proceed to the analysis of the Principal's equilibrium offer. First, I prove the monotonicity Lemma, which is based on a standard revealed preferences argument.

Lemma 6. *For any $q_L^* \in \mathrm{supp}\ \sigma_L^*$, $q_H^* \in \mathrm{supp}\ \sigma_H^*$ holds*

$$C(q_H^*)a^*(q_H^*) \le C(q_L^*)a^*(q_L^*)$$
$$q_H^* \le q_L^*$$

The Proof of Lemma 6 is given in the Appendix.

Next, I show that if an equilibrium has a pooling part, it can consist of only one offer.

Lemma 7. *If* supp $\sigma_H^* \cap$ supp $\sigma_L^* \neq \varnothing$ *then there is only one common point in the supports of the equilibrium mixed strategies for the two types:*

$$\text{supp } \sigma_H^* \cap \text{supp } \sigma_L^* = \left\{q_p^*\right\}$$

The Proof of Lemma 7 is given in the Appendix.

The two Lemmas 6 and 7 show that any PBE has a very particular structure: there can be a pooling part - an offer q_p^* made by both types of the Principal, and a separating part - the offers made only by α_H-type lying to the left of q_p^*, and the offers made only by α_L-type to the right of q_p^*. Put formally,

$$q_H^* < q_p^* < q_L^* \tag{28}$$

for any $q_H^* \in$ supp $\sigma_H^* \setminus$ supp σ_L^*, $q_L^* \in$ supp $\sigma_L^* \setminus$ supp σ_H^* if the corresponding elements of an equilibrium exist and given that $a^*(q_H^*) = 1$.

I will distinguish between

- pooling equilibria - equilibria with supp $\sigma_H^* =$ supp σ_L^*,
- semi-separating equilibria[23] - equilibria with supp $\sigma_H^* \neq$ supp σ_L^* and supp $\sigma_H^* \cap$ supp $\sigma_L^* \neq \varnothing$
- separating equilibria, for which supp $\sigma_H^* \cap$ supp $\sigma_L^* = \varnothing$.

For the further analysis the relative position of $q_H^P, q_L^P, q_L^0, q_E^0$ is important. It is partially described by (23) and (24). More precise characterization can be obtained by inspecting Figure 6. The formal statement is as follows.

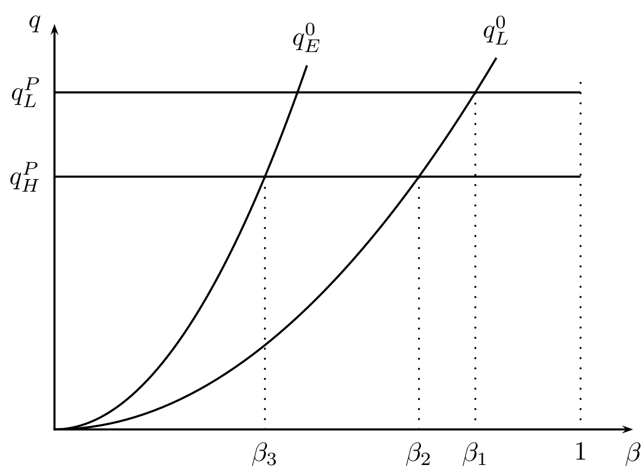

Figure 6: The relative position of $q_H^P, q_L^P, q_L^0, q_E^0$

[23]Semi-separating equilibria are often called "hybrid" in the literature.

Proposition 4. *There exist thresholds $\beta_3 < \beta_2 < \beta_1$ determined by*

$$\begin{array}{lll} \beta_1 & \textit{is solution to} & q^0(\alpha_L, \beta) = q_L^P \\ \beta_2 & \textit{is solution to} & q^0(\alpha_L, \beta) = q_H^P \\ \beta_3 & \textit{is solution to} & q^0(E\alpha, \beta) = q_H^P \end{array}$$

such that

$$\begin{array}{ll} q_L^0 < q_H^P & \textit{for } \beta < \beta_2 \\ q_H^P < q_L^0 & \textit{for } \beta > \beta_2 \end{array}$$

The thresholds β_j are important to describe the equilibria structure and properties.

For the analysis of the separating and semi-separating equilibria the crucial is the relative position of q_H^P and q_L^0, i.e. whether $\beta > \beta_2$ or $\beta < \beta_2$.

The following lemma characterizes the set of the PBE in the signaling game.

Lemma 8. *1. For $\beta \leq \beta_2$ the signaling game has multiple pooling and semi-separating PBE.*

In any pooling equilibria there is only one offer q_p^* made by both Principal types: $\text{supp } \sigma_H^* = \text{supp } \sigma_L^* = \{q_p^*\}$.

The offer q_p^* is a pooling equilibrium offer iff $q_L^0 \leq q_p^* \leq \min\{q_E^0, q_L^{00}\}$ where $q_L^{00} > q_L^0$ is the solution to $V(q_L^{00}, \alpha_H) = V(q_L^0, \alpha_H)$.

In any semi-separating equilibrium there can be one or two offers $q_H^*(q_H^{**}) \in$ $\text{supp } \sigma_H^* \setminus \text{supp } \sigma_L^*$, made by α_H-type only, and one pooling offer q_p^*, with the acceptance probabilities $a^*(q_H^*) = 1$, $a^*(q_p^*) \leq 1$; for $\beta \leq \beta_3$ there is only one offer $q_H^* \in \text{supp } \sigma_H^* \setminus \text{supp } \sigma_L^*$.

2. For $\beta > \beta_2$ the signaling game has only unique separating PBE. In this equilibrium $q_H^* = q_H^P$, $q_L^* = \min\{q_L^0, q_L^P\}$ and $q_L^* > q_H^*$.

The Proof of Lemma 8 is given in the Appendix.

Now we have obtained the structure of the set of PBE of the signaling game in great details.

The multiplicity of equilibria is the consequence of no restrictions on the acceptance rule for the out-of-equilibrium offers. Equilibrium refinement is a standard procedure for the signaling games which allows to eliminate "unreasonable" equilibria.

4.1.2 Equilibrium Refinement

I will now focus on the case $\beta \leq \beta_2$, since there are multiple PBE in this case, whereas for $\beta > \beta_2$ there is unique PBE[24]. I use the intuitive criterion[25] proposed by Cho and Kreps [1987]. I'll show that it eliminates some equilibria for the case $\beta \leq \beta_2$, though there is still a continuum of equilibria satisfying it. I will argue then that one of them is "the most reasonable" by applying a stronger refinement.

Some notation is needed to implement the intuitive criterion, following Cho and Kreps [1987]. Let $\nu(\mathcal{A}'|q)$ be the ex-post probability which the Agent assigns

[24] It's easy to show that the separating equilibrium satisfies the intuitive criterion.

[25] A survey of the refinements procedures and approaches can be found in Fudenberg and Tirole [1991].

to the (sub)set $\mathcal{A}'$ of the Principal's type after observing an offer q. For the two-type case $\mathcal{A}'$ can be $\{\alpha_H\}$, $\{\alpha_L\}$ or $\mathcal{A} = \{\alpha_H, \alpha_L\}$.

Formally,

$$\nu(\mathcal{A}'|q) = \mu(q) I_{\{\alpha_H \in \mathcal{A}'\}} + (1 - \mu(q)) I_{\{\alpha_L \in \mathcal{A}'\}}$$

Let

$$BR(\mathcal{A}', q) = \bigcup_{\mu:\ \nu(\mathcal{A}'|q)=1} BR_\mu(q)$$

be the set of all reasonable acceptance rules for beliefs concentrated on the (sub)set $\mathcal{A}'$ applied to the offer q would it be proposed. The acceptance rule is "reasonable" if it is a Best Response corresponding to some beliefs concentrated on the (sub)set $\mathcal{A}'$.

Fix an equilibrium profile $((\sigma_H^*(\cdot), \sigma_L^*(\cdot)), a^*(\cdot); \mu^*(\cdot))$. Denote the equilibrium payoff of α_j-type Principal by $V_j^* = V(q_j^*, a^*(q_j^*), \alpha_j)$ for some[26] $q_j^* \in \text{supp } \sigma_j^*$.

Let $J(q)$ be the set of types which for sure don't want to deviate to q for any reasonable acceptance rule:

$$J(q) = \left\{ \alpha_j :\ V_j^* > \max_{a \in BR(\mathcal{A}, q)} V(q, a, \alpha_j) \right\}$$

The set $\mathcal{A} \setminus J(q)$ then is the set of Principal's types which for sure want to deviate from an equilibrium and make an offer q, provided that some reasonable acceptance rule will be applied.

The equilibrium satisfies the intuitive criterion if

$$V_j^* \geq \min_{a \in BR(\mathcal{A} \setminus J(q), q)} V(q, a, \alpha_j) \quad \text{for all } j, q \tag{29}$$

The following Proposition states the main result of the equilibrium refinement of the signaling game.

Proposition 5. *For $\beta \leq \beta_2$ $\left(\Leftrightarrow\ q_L^0 \leq q_H^P\right)$ only the pooling equilibria with*

$$q_L^0 \leq q_p^* \leq q_E^0$$

satisfy the intuitive criterion.

The Proof of Proposition 5 is given in the Appendix.

For the case of $\beta \leq \beta_2$ $\left(\Leftrightarrow q_L^0 \leq q_H^P\right)$ there are still many pooling equilibria, satisfying the intuitive criterion, though fewer than in the set of PBE and none of the semiseparating equilibria don't pass the intuitive criterion. It can be shown that the pooling equilibria, satisfying the intuitive criterion, can't be eliminated by applying Criterion D1 (a version of the divinity equilibrium in the sense of Fudenberg and Tirole [1991], ch.11) or NWBR criteria.

The reason for the multiplicity of the pooling equilibria, satisfying the intuitive criterion is that the equilibrium payoff for the Principal is compared in (29) with the worst (for the Principal) reasonable acceptance rule, based on Agent's beliefs, concentrated on the set $\mathcal{A} = \{\alpha_H, \alpha_L\}$. In the worst case, after observing a deviation to $q > q_p^*$, the Agent believes that this deviation is done

[26] Clearly, for any $q_j^* \in \text{supp } \sigma_j^*$, $V(q_j^*, a^*(q_j^*), \alpha_j)$ takes the same value, so the definition of V_j^* is correct.

by α_L-type. Then, any offer $q > q_L^0$ is reasonably rejected and the intuitive criterion (29) is satisfied for PBE offers $q_p^* > q_L^0$, since any upward deviation from q_p^* is reasonably rejected.

However, once such upward deviation is profitable for both types, the intuitive criterion can be strengthened by comparing the equilibrium payoff in (29) with payoff obtained under "more reasonable" acceptance rule which is based on the ex-ante beliefs instead of the worst beliefs. Then, in the right-hand side of (29) the acceptance rule applied to deviations $q > q_{HL}^*$ is the Best Response acceptance rule based on beliefs $\mu(q) = \Pi$, so that any $q < q_E^0$ (not only $q < q_L^0$) is accepted. This rules out all the PBE with $q < q_E^0$.

For an equilibrium offer $q_p^* = q_E^0$ the Agent is indifferent between accepting and rejecting (and between any probability of accepting), but for the acceptance rule satisfying the strengthened intuitive criterion, all the offers just below q_E^0 are accepted with probability 1, so if $a^*(q_E^0) < 1$, then there will be profitable deviation for any Principal to $q_E^0 - \varepsilon$. Consequently, only $a^*(q_E^0) = 1$ is possible in a (refined) equilibrium.

The strengthening of the intuitive criterion in this way is equivalent to requiring the acceptance rule to be sequentially rational on the out-of-equilibrium path. It is also equivalent to eliminating the weakly dominated acceptance rules.

Finally, the "non-worth beliefs" intuitive criterion selects in the considered game the Pareto-dominant pooling equilibrium, which is the equilibrium with the highest effort.

As a result, only equilibrium with the offer $q_p^* = q_E^0$ for $\beta \leq \beta_3$ satisfies the strengthened intuitive criterion. For $\beta_3 < \beta < \beta_2$ the strengthened criterion also leads to the unique prediction for the game outcome $q_p^* = q_H^P$. Indeed, this offer is feasible, and α_H-type prefers to make it. By deviating to a higher offer, α_L-type would be revealed and then the deviating offer would be rejected, so α_L-type has to pool on q_H^P.

It's easy to check that the unique separating equilibrium for $\beta > \beta_2$ which satisfies the intuitive criterion satisfies the strengthened criterion as well.

4.2 The 2-type Signaling Game Outcome

For the rest of the discussion I call α_L-type the "tough" Principal and α_H-type - the "generous" Principal. Consider an intuitive explanation for the signaling game outcome.

When the Agent is highly reciprocal, $\beta \in (\beta_1, 1]$, he can be easily intrinsically motivated even by the tough Principal and agrees to perform at her preferred level q_L^P, which is quite high, because the Agent's participation constraint isn't very tight. In this case, the tough Principal will reveal her type by requiring performance at this high level.

However, as the Agent's reciprocity declines, his intrinsic motivation decreases, and the tough Principal can't inspire the Agent to perform at the level q_L^P and has to follow the Agent participation constraint by offering contract with $q < q_L^P$, still revealing her type for $\beta \in (\beta_2, \beta_1)$. The Agent gets zero social utility in this case.

For even lower reciprocity intensity, the Agent's intrinsic motivation is not enough to make him exert effort higher than q_H^P if he would learn that the Principal is tough. In this case, the tough Principal follows the offer of the generous Principal, in other words, she has to mimic the generous Principal.

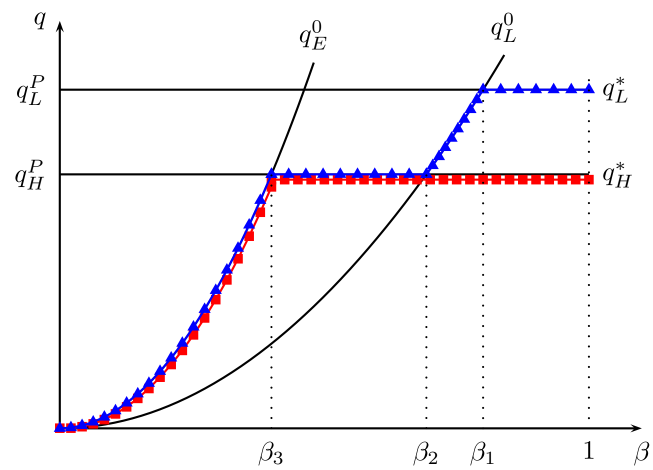

Figure 7: Equilibrium in the signaling game

The Agent can't distinguish the two types of the Principal and his participation threshold becomes being based on the expected value of the Principal altruism $E\alpha > \alpha_L$ so that $q = q_H^P$ breaks even. The generous Principal doesn't want to separate since she gets her preferred output q_H^P. The Agent's expected utility is positive. This is the case for $\beta \in (\beta_3, \beta_2)$.

Finally, when the Agent's reciprocity intensity is very low so that the Agent doesn't have enough intrinsic motivation to exert effort q_H^P, both types of Principal follow the Agent's participation threshold for the expected value of the Principal's altruism $E\alpha$, not revealing the type. Neither type of the Principal has an incentive to deviate and reveal his type. This is the case for $\beta \in (0, \beta_3)$.

To conclude the description of the outcome, I stress that in the 2-type signaling game the tough Principal doesn't want to mimic the generous one unless she has to, because mimicking will result in lower output which is not desirable for the tough Principal. However, the Agent is (intrinsically) motivated to accept an offer if it isn't too high. As a result, if the Agent's intrinsic motivation isn't high enough - due to low intensity of reciprocity or due to revealed low altruism of the tough Principal, the tough Principal has to follow the Agent's participation threshold or the generous Principal's offer.

4.3 Signaling with Continuum of Types

The analysis for 2 types can be generalized to the case of a continuum of types. I don't present the complete analysis, as for the 2-type case and rather focus on the "most reasonable" equilibrium[27], which leads to the unique prediction of the signaling game outcome.

Let the Principal's altruism parameter α be distributed on the interval $[\alpha_1, \alpha_2] \subset [0, 1]$ with continuous CDF $F(\alpha)$.

The interval's bounds α_1 and α_2 are the exact bounds of the distribution:

[27] See the refinement section for the 2-type case for the discussion of the "most reasonable" equilibrium.

$\alpha_1 = \inf\{\alpha|F(\alpha) > 0\}$, $\alpha_2 = \sup\{\alpha|F(\alpha) < 1\}$.

Let $\alpha^{\times}$ be the solution to

$$q^0\left(\alpha^{\times},\beta\right) = q^P\left(\alpha^{\times}\right)$$

which always exists[28] and is unique since the left-hand side is increasing and the right-hand side is decreasing function of α - see Figure 8.

The next property follows directly from the definition of $\alpha^{\times}$.

Claim 4. *The Principal's preferred output q^P is feasible, i.e. satisfies the Agent's participation constraint iff $\alpha \geq \alpha^{\times}$.*

So, the population of the Principals can be separated into two sub-populations. One consists of comparatively generous ones with $\alpha \geq \alpha^{\times}$, which can inspire the Agent's intrinsic motivation high enough to implement their preferred output q^P, would their altruism be revealed. Another subpopulation consists of comparatively tough Principals with $\alpha < \alpha^{\times}$, which can not inspire high enough Agent's intrinsic motivation.

Denote by

$$E_{\widetilde{\alpha}}[\alpha] = E[\alpha|\alpha < \widetilde{\alpha}]$$

the expected value of the truncated distribution of Principal's altruism parameter, bounded at the top by $\widetilde{\alpha}$.

The signaling game outcome in the "most reasonable" equilibrium is characterized by the following Proposition.

Proposition 6. *1. If α is distributed inside the interval $[\alpha^{\times}, 1]$, then all the Principals implement their preferred output in the "most reasonable" equilibrium, i.e. the equilibrium contract is*

$$q = q^P(\alpha)$$

2. If $\alpha_1 < \alpha^{\times}$ and

$$q^0\left(E\alpha,\beta\right) > q^P\left(\alpha_2\right) \tag{30}$$

then there exists $\tilde{\alpha} \in [\alpha^{\times}, \alpha_2]$ determined as solution to

$$\widetilde{q}^0 \equiv q^0\left(E_{\widetilde{\alpha}}[\alpha],\beta\right) = q^P\left(E_{\widetilde{\alpha}}[\alpha]\right) \tag{31}$$

such that the "most reasonable" equilibrium contract is given by

$$q = \begin{cases} q^P(\alpha,\beta) & \text{for } \alpha > \widetilde{\alpha} \\ \widetilde{q}^0 & \text{for } \alpha \leq \widetilde{\alpha} \end{cases}$$

where $\widetilde{q}^0$ is determined by (31).

3. If the inequality (30) doesn't hold[29], then the "most reasonable" equilibrium contract is full pooling with

$$q = q^0\left(E\alpha,\beta\right)$$

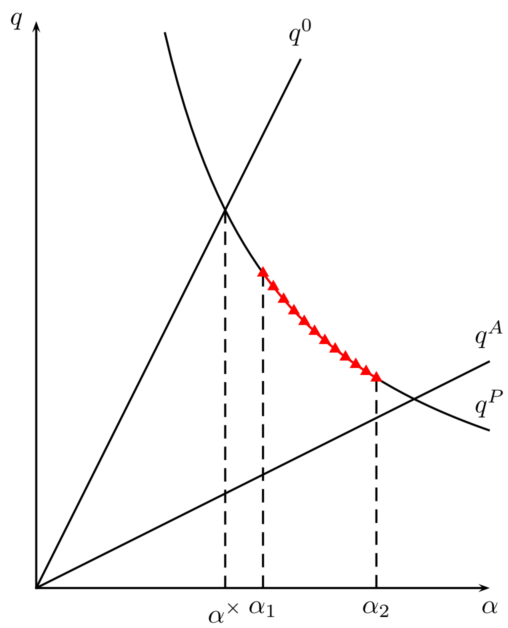

Figure 8: Equilibrium in signaling game. Case 1

The Proof of Claim Proposition 6 is given in the Appendix.

Figure 8 illustrates the equilibrium contract for the case when ex-ante all the Principals in the population are highly-altruistic, i.e. $\alpha \geq \alpha^{\times}$ and can inspire the Agent to perform at their preferred level (point 1 of the Proposition).

Figure 9 illustrates the case when there are both generous and tough Principals, i.e. with altruism greater and smaller than $\alpha^{\times}$, but there are enough Principals with high altruism (point 2 of the proposition). In this case only a part of the Principals' subpopulation with $\alpha > \alpha^{\times}$, for instance those with $\alpha \geq \widehat{\alpha}$ implement their preferred performance level q^P.

The first case from Proposition 6 emerges for high β. As β decreases, the equilibrium structure switches to the one described at point 2 of the Proposition and then to the one of point 3.

Similarly to the 2-type case, when the Agent is highly reciprocal, the separating equilibrium in which all the Principals implement their preferred output emerges. For lower levels of reciprocity, the equilibrium structure shifts to pooling and a larger share of Principals can't implement their preferred performance level.

4.4 Application to the Organization Design

For the 2-type framework, the generous Principal implements her preferred output for $\beta \in [\beta_3, 1]$. In this case she cares neither about the altruism level of the tough Principals α_L, nor about the structure of the Principals' population characterized by Π.

[28]The fact that it can be that $\alpha^{\times} > 1$ is not a problem.

[29]This implies $\alpha_1 < \alpha^{\times}$.

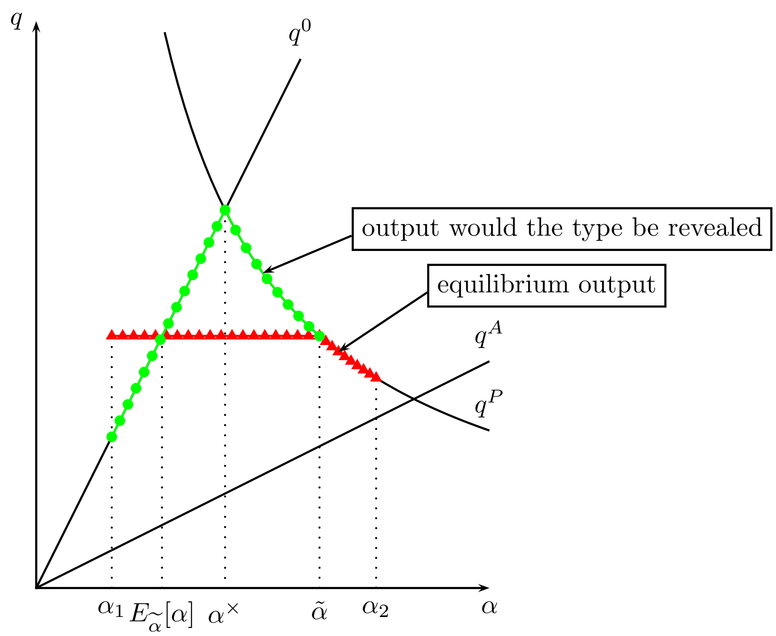

Figure 9: Equilibrium in signaling game. Case 2

However, if $\beta \in [0, \beta_3)$, the generous Principal can't implement her preferred output in the emerging pooling equilibrium. In this case she is affected by the adverse effect which emerges due to the very existence of the tough Principals in the population.

Such adverse effect is present for the continuum-type framework as well. In particular, for the case illustrated by Figure 9, the Principals with $\alpha \in [\alpha^{\times}, \tilde{\alpha})$ are affected. For the case, described in point 3 of Proposition 6, the whole generous subpopulation with $\alpha \in [\alpha^{\times}, \alpha_2]$ is affected.

Even a part of the "tough" Principals with $\alpha < \alpha^{\times}$ is affected as they get output lower than the Agent's participation threshold would the Principals' type be revealed. For the case depicted by Figure 9, these have $\alpha \in (E_{\widehat{\alpha}}[\alpha], \alpha^{\times}]$; for the case of point 3 of Proposition 6 these have $\alpha \in (E\alpha, \alpha^{\times}]$.

In both frameworks - the 2-type and the continuum-type,

- the most altruistic Principals (among the generous ones) are less likely to be affected,
- if Agent's reciprocity is high, then all the concerned Principals are less likely to be affected,
- if tough Principals are not too tough or if they represent a smaller share in the population (so that the expected altruism level in the whole population is high enough), then all the concerned Principals are less likely to be affected.

This means that the most generous Principals care to a lesser extent about the environment in which they operate. For the Principals which are generous but not the most generous ones, the environment can play an important role for their ability to elicit a high performance level and they prefer to be in the environment of the generous Principals and highly reciprocal Agents.

The counterpart of the adverse effect is an advantageous effect for the tough Principals which emerges due to the very existence of the generous Principals in the population. In the 2-type framework, for the case of $\beta \in [0, \beta_2)$ the tough Principal gets an output which exceeds the Agent's participation threshold would the type of the tough Principals be revealed. In the continuum-type framework, some Principals benefit from the same advantage: for the case depicted by Figure 9, these are ones with $\alpha \in [\alpha_1, E_{\widetilde{\alpha}}[\alpha])$. The advantageous effect appears when the tough Principal can mimic to be the generous ones (or when these more generous can't separate from the tougher ones).

As a result, the tough Principals prefer having highly reciprocal Agents and if this is not the case - then to be in the environment of the generous Principals to benefit from the advantageous effect in the pooling equilibrium.

Now let us take a point of view of an organization designer. Assume that he has a choice - whether to create an organizations of type A in which the Principals' population is mixed, or create an organization of type B in which the generous and tough Principals are separated. The organization designer informs in this way the Agents about the Principal's altruism (the information can still be imprecise).

For instance, in the 2-type framework, the organization designer can separate the two types, in the continuum-type framework, the designer can separate the Principals with $\alpha > \alpha^{\times}$ from those below the threshold $\alpha^{\times}$.

An interesting question is how the choice of organization type influences the overall performance in the organization?

Clearly, in the organizations of type A, the generous Principals will implement a lower output, whereas in the organizations of type B the generous Principals will get a higher output. This shows that if separation emerges as a result of signaling (endogenously), the tough Principals implement a higher output; if separation is exogenous, then the relation between the outputs, implemented by the tough and generous Principals, is reversed.

Figure 10 illustrates exogenous Principals' separation compared to the endogenous Principals' population for the continuum-type framework.

4.5 A "Unifying" framework

The three models considered above - the model without extrinsic incentives and the two models related to the experiments have a lot in common and can be unified in the following way.

At stage 1, the Principal sends a message - fixes $\widehat{Q}$ - the set of feasible performance levels for the Agent. Performance is costly for the Agent. At stage 2, the Agent decides whether to comply or disobey. In case of compliance, the Agent chooses an element $q \in \widehat{Q}$ - the implemented performance level. In case of disobeying, the Agent chooses a disobedience option in a set of disobedience options[30] $\mathring{Q}$: $\mathring{q} \in \mathring{Q}$.

[30]The set of performance levels can be considered as a subset of the disobedience options: $\widehat{Q} \subset \mathring{Q}$. Alternatively, the set of disobedience options can be considered as a set of all possible Agent's actions.

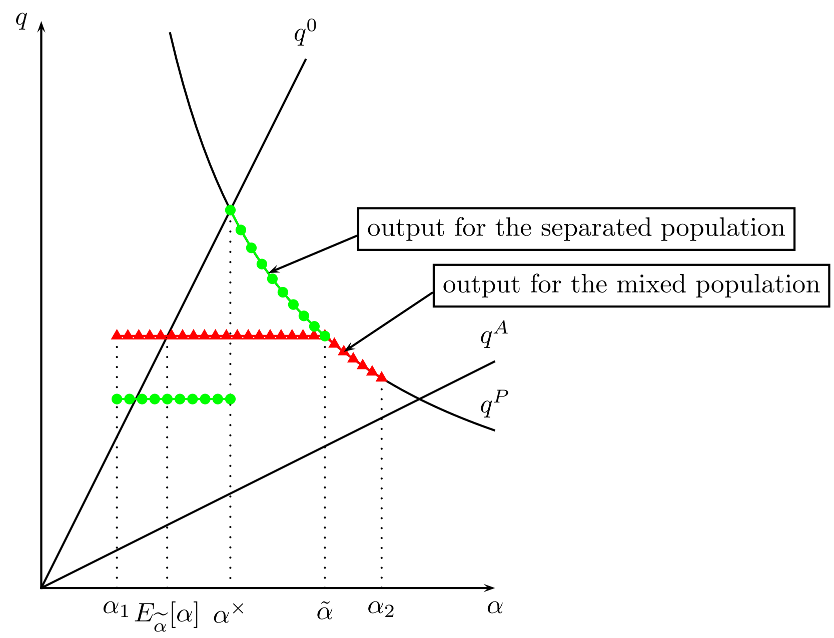

Figure 10: Application to the organization design. The effect of exogenous separation of the Principals' population

For the model without extrinsic incentives $\widehat{Q} = \{q\}$ - the required output level, compliance corresponds to agreement to perform at level q ($a(q) = 1$), disobeying corresponds to $\mathring{q} = 0$ or, equivalently, to $a(q) = 0$. For the Trust Game the performance set is $\widehat{Q} = [\widehat{q}, +\infty)$, compliance corresponds to the choice of some back-transfer $q \in [\widehat{q}, +\infty)$, disobeying corresponds to the choice of $\mathring{q} < \widehat{q}$, in the equilibrium with disobedience $\mathring{q} = q^A$. Finally, for the Control game $\widehat{Q} = (q_c, +\infty)$, compliance means the choice of effort higher than the minimal requirement, disobeying is the choice of effort at the minimal level: $\mathring{q} = q_c$.

All the three models are based on the reciprocal altruism framework. This means that the Agent is intrinsically motivated but there is a gap between the Principal's preferred performance level and that of the Agent. The Agent prefers the lower performance.

The weaker the Principal's altruism (i.e. the less generous she is), the larger is the gap, the more the Principal suffers from leaving the Agent too much flexibility, i.e. not restricting at all or imposing only slight restrictions on the performance set $\widehat{Q}$. Consequently, it's in the Principal's interest (especially, the tough Principal) to make a strict offer - i.e. to restrict the Agent's performance set $\widehat{Q}$ to avoid a large deviation from her preferred performance level. On the other hand, offering a slightly restricted or non restricted at all, compared to $\mathring{Q}$, performance set $\widehat{Q}$ is a generous offer, which is likely to be made by the highly altruistic Principal.

By offering a generous performance set $\widehat{Q}$, the Principal can signal her generosity. By learning the Principal's generosity, the Agent becomes inspired for better performance, i.e. intrinsically motivated. An extrinsic motivator, added to the intrinsic motivation, leads to even better performance for the symmetric information on the Principal's generosity. However, for the case of asymmetric information, the use of an extrinsic motivator signals the non-generosity of the Principal (for the case of separating equilibrium) and crowds out the Agent's intrinsic motivation.

Such a crowding mechanism influences the Agent's performance for the Trust Game and for the Control game - performance for the tough Principal can be lower than for the generous one, despite the low performance is feasible for the performance set, proposed by the highly altruistic Principal.

The motivation crowding out is presented in the model without extrinsic incentives as well, though it doesn't result in the crowding-out of performance, because the signaling of generosity goes through offering the performance level itself, not the (non-trivial) set of the performance levels, as in the Trust and Control games.

In the model without extrinsic incentives, the Agent attached to the generous Principal gets higher utility from an interaction. Indeed, he is intrinsically motivated for high performance (the participation threshold is high) but is asked for lower performance, and, consequently, gets positive utility. The Agent attached to the tough Principal gets positive utility only for the separating equilibria with $\beta \in (\beta_1, 1]$ - see Figure 7. For $\beta \in (\beta_2, \beta_1]$ the Agent performs at his participation threshold and gets zero utility, for $\beta \leq \beta_2$ the Agent gets negative utility in the pooling equilibria.

Another common feature of the Trust and the Control games is the effect of the strength of an extrinsic motivator. If it is weak, not using it can not signal the Principal's generosity. In fact, in the Control Game if the threshold q_c is small, it is not used by both types[31] and has no effect at the performance level. In the Trust Game, if the fine $\overline{f}$ is small, it is used by both types (the two lines q^A and $\widehat{q}^A$ are close to each other so that pooling with both types imposing the fine emerges).[32]

If the available extrinsic motivator is strong enough but not too strong and Agent's reciprocity intensity is high enough, the generous Principal can signal her generosity by not using the extrinsic motivator. The tough Principal prefers to reveal her type and to crowd out the Agent's intrinsic motivation in this way because she gets better performance by using the extrinsic motivator and compensating in this way the crowding out of the intrinsic motivation.

Finally, if the available extrinsic motivator is very strong, both types can benefit form using it. The generous Principal crowds out the intrinsic motivation to the pooling level but gets compensated through high additional extrinsic motivation of the Agent.

5 Conclusion

My paper contributes both to the relatively new field of behavioral economics and to the theory of incentives. More specifically, I develop a model which goes

[31] Alternatively, it can be used by both types - see case I of proposition 2.

[32] Only the Control game experiment provides an evidence supporting the reciprocal altruism model. There were no variation of the size of the extrinsic incentive, i.e. the fine, in the Trust Game experiment.

beyond the neoclassical framework and takes into account the social components of preferences. I build a simple and intuitive model of reciprocal altruism and show its relevance for the contract theory and theory of incentives as it accounts for the behavioral patterns, systematically observed both in the lab and in the field.

The relevance of altruism and reciprocity as parts of human nature is supported by compelling evidence. The Dictator Game introduced in Kahneman et al. [1986] provides evidence for pure altruism[33]. The survey of Andreoni [2006] demonstrates that impure altruism - taste for "warm glow" shapes people's decisions in many circumstances. The evidence for reciprocity[34] is provided by Berg et al. [1995] who introduced the Trust Game, which was repeated with modifications in Fehr and Rockenbach [2003], Fehr and List [2004] and others. The evidence also comes from variants of the Gift Exchange Game by Fehr et al. [1997]; Lost Wallet Game by Dufwenberg and Gneezy [2000] and Charness et al. [2007]; Moonlighting Game by Abbink et al. [2000].

The model of the paper predicts that crowding-out in performance can emerge as the equilibrium outcome for some values of parameters of the model. The crowding-out in performance is the situation in which imposing an extrinsic incentive decreases the intrinsic motivation and leads to a lower performance - the phenomena, well-known in human resources management. It can also be the case (for some other values of parameters) that crowding-out in incentives doesn't result in crowding-out in performance.

I show that for the crowding-out equilibrium to emerge, the sorting condition is required, which, in turn is guaranteed when the available extrinsic incentives is neither too weak neither too strong.

This paper is among a few others studies aimed at enriching the theory of incentives by taking into account intrinsic motivation. Further research can be devoted to considering other information structures - Agents can differ in productivity, the organization can have more complicated structure than only one Principal and only one Agent. Considering other social components of preferences, supported by the evidence from lab and field (negative reciprocity, concerns for equity etc.) can also be relevant in building the theory of incentives.

Finally, I don't claim that monetary incentives are not important. On the contrary, it is well-known that the incentive payments play an important role in creating incentives - see e.g. Bolton and Dewatripont [2005] or Prendergast [1999]. However, there is growing evidence that workers are motivated to exert effort not only by the incentive payment or other extrinsic motivators but also by the intrinsic motivation. On top of this, the interaction between intrinsic and extrinsic motivation can play an important role. The result of such interaction can be motivation crowding-out (or -in). Therefore, taking into account the intrinsic motivation in the labor contract models should give better understanding of the workplace relation. Intrinsic motivation and extrinsic incentives should be considered as complements rather than substitutes in the modeling.

[33]In the various Dictator Game experiments, the subjects are endowed with a sum of money. They decide then on how much of this windfall endowment to give to a stranger. More than half of the subjects give between 20% and 50% of the endowment.

[34]Precisely, "intrinsic reciprocity", not "consequentialism" or "strategic reciprocity".

6 Appendix

Proof of Lemma 4

Proof. The root exists and is unique since $U(0) = B > 0$, $U(q)$ increases for $q \in (0, q^A)$, so that $U(q^A) > 0$, then decreases for $q \in (q^A, \infty)$ and $U(q) \to -\infty$ as $q \to \infty$. Because of continuity of $U(q)$, there exists a unique $q^0 \in (q^A, \infty)$ such that $U(q^0) = 0$. □

Proof of Claim 1

Proof. Statements 1 is trivial since the Agent has full flexibility and hence chooses his preferred back-transfer.

For the 2-nd point, notice that if the desired back-transfer $\widehat{q} \leq q^A$, then paying back q^A will not impose fine and will maximize the Agent's utility (q^A is the global maximizer).

Consider the case $\widehat{q} > q^A$.

Notice that $\widetilde{q}^A(\widehat{\alpha})$ is constructed in such way that

$$\overset{\circ}{U}(q) > \overset{\circ}{U}(q^A) - f \quad \text{for} \quad q^A(\widehat{\alpha}, \beta) < \widehat{q} < \widetilde{q}^A(\widehat{\alpha}, \beta) \tag{32}$$

$$\overset{\circ}{U}(q) < \overset{\circ}{U}(q^A) - f \quad \text{for} \quad \widehat{q} > \widetilde{q}^A(\widehat{\alpha}, \beta) \tag{33}$$

where $\overset{\circ}{U}(q)$ is the Agent's utility without taking into account the possibility of fine: $U(q) = \overset{\circ}{U}(q) - f I_{q<\widehat{q}}$.

In (32) the Agent prefers to diverge from q^A to $q > q^A$ as such divergence isn't too high whereas in (33) the Agents prefers to pay fine. □

Proof of the Proposition 1

Proof. The optimality of the Principal's decision given the Agents's beliefs is evident from Claim 1. We should check the incentives compatibility conditions and the crowding-out condition.

Consider the case $q_{LH} \leq \widetilde{q}_{LL}$.

The Principal's incentive compatibility constraints are

$$\pi_H(q_{HH} - \alpha_H C(q_{HH})) + (1 - \pi_H) \cdot 0 \geq \widetilde{q}_{LL} - \alpha_H C(\widetilde{q}_{LL})$$
$$\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL}) \geq \pi_L(q_{HH} - \alpha_L C(q_{HH})) + (1 - \pi_L) \cdot 0$$

which are equivalent to (12) and (13), correspondingly.

The inequality $\widehat{\pi}_H \leq 1$ holds since the denominator in (12) is positive and the inequality is then equivalent to

$$\frac{C(q_{HH}) - C(\widetilde{q}_{LL})}{q_{HH} - \widetilde{q}_{LL}} \leq \frac{1}{\alpha_H}$$

The left-hand side $\frac{1}{\alpha_H} \geq 1$. The right-hand side is the slope of the secant line to the graph of the convex function $C(q)$ between the points with $q = \widetilde{q}_{LL}$ and $q = q_{HH}$, which is smaller than the slope of the tangent line at the point with $q = q_{HH}$ which is equal to $C'(q_{HH}) = \alpha_H \beta_H < 1$. So, the inequality holds.

The inequality $\widehat{\pi}_L > 0$ holds since both the numerator and the denominator of the fraction are positive.

Finally, we need to check the crowding-out condition $\pi q_{HH} \geq \widetilde{q}_{LL}$, where π is the objective probability of the selfish Agents. Since the selfish Principal has beliefs π_L which is biased downward, it is sufficient to prove that $\widehat{\pi}_L q_{HH} \geq \widetilde{q}_{LL}$. Substituting $\widehat{\pi}_L$ into the inequality, we get

$$\frac{\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})}{q_{HH} - \alpha_L C(q_{HH})} q_{HH} \geq \widetilde{q}_{LL}$$

Since the denominator is positive, this inequality is equivalent to

$$\alpha_L \widetilde{q}_{LL} q_{HH} \left(\frac{C(q_{HH})}{q_{HH}} - \frac{C(\widetilde{q}_{LL})}{\widetilde{q}_{LL}} \right) \geq 0$$

which holds since $q_{HH} > \widetilde{q}_{LL}$, $\widetilde{q}_{LL} < \widetilde{q}_{LH}$ and it's assumed that $\widetilde{q}_{LH} < q_{HH}$. $\square$

Proof of the Corollary 1

Proof. The condition $q_{LH} \leq \widetilde{q}_{LL}$ is equivalent to $C(q_{LH}) \leq C(\widetilde{q}_{LL})$. Since $C(\widetilde{q}_{LL}) = f$, it leads to $C(q_{LH}) \leq f$, so that $f_1 = C(q_{LH})$.

Now check the condition $\widetilde{q}_{LH} \leq q_{HH}$.

The back-transfer $\widetilde{q}_{LH}$ is determined, according to Claim 1 by

$$\alpha_L \beta_H q_{LH} - C(q_{LH}) - f = \alpha_L \beta_H \widetilde{q}_{LH} - C(\widetilde{q}_{LH})$$

where $\widetilde{q}_{LH}$ is chosen on the decreasing part of the graph of the function $F(q) = \alpha_L \beta_H q - C(q)$ (see Figure 2). Consequently, $\widetilde{q}_{LH} \leq q_{HH}$ is equivalent to

$$\alpha_L \beta_H q_{LH} - C(q_{LH}) - f \geq \alpha_L \beta_H q_{HH} - C(q_{HH})$$

which can be rewritten as

$$f_2 \equiv (\alpha_L \beta_H q_{LH} - C(q_{LH})) - (\alpha_L \beta_H q_{HH} - C(q_{HH})) = f_2 \geq f$$

Finally, to make sure that the interval $[f_1, f_2]$ is non-empty, we should check that $f_1 \leq f_2$. This leads to

$$\alpha_L \alpha_H \beta_H^2 \left(\frac{q_{LH}}{\alpha_L \beta_H} - \frac{q_{HH}}{\alpha_H \beta_H} \right) \leq C(q_{LH})$$

for a generic cost function.

For the quadratic cost function $C(q) = \frac{c}{2} q^2$, taking into account that q_{ij} are determined by $C'(q_{ij}) = \alpha_i \beta_j$, and substituting this into the last inequality, one can check that the left-hand side is equal to zero, so that the inequality always holds.

Finally, for given $\alpha_L, \alpha_H, \beta_H$, and $f \in [f_1, f_2]$, one can obtain the threshold values $\widehat{\pi}_L \geq 0$, $\widehat{\pi}_H \leq 1$ from (12) and (13), and take the values π_L and π_H, satisfying $\pi_H \geq \widehat{\pi}_H$, $\pi_L \leq \widehat{\pi}_L$. For these parameters, according to Proposition 1, the equilibrium of the signaling game is the separating crowding-out equilibrium. $\square$

Proof of Proposition 2

Proof. For the case $q_c \geq q_{LH}$, the incentives compatibility constraints for the Principal are:

$$\pi_H(q_{HH} - \alpha_H C(q_{HH})) \geq q_c - \alpha_H C(q_c)$$
$$q_c - \alpha_L C(q_c) \geq \pi_L(q_{HH} - \alpha_L C(q_{HH}))$$

These constraints are equivalent to the conditions $\pi_L \leq \widehat{\pi}_L$ and $\pi_H \geq \widehat{\pi}_H$, which are assumed to hold.

Check the crowding-out condition $\pi q_{HH} \geq q_c$. Since $\pi > \pi_L$, the inequality $\pi_L q_{HH} \geq q_c$ is stronger. I check the latter inequality for $\pi_L = \widehat{\pi}_L$.

Substituting the formulae for $\widehat{\pi}_L$, we get

$$\frac{\widetilde{q} - \alpha_L C(\widetilde{q})}{q_{HH} - \alpha_L C(q_{HH})} q_{HH} \geq \widetilde{q}$$

after rearranging it leads to

$$\frac{C(\widetilde{q})}{\widetilde{q}} \leq \frac{C(q_{HH})}{q_{HH}}$$

which is equivalent to $\widetilde{q} \leq q_{HH}$ since the function $C(q)$ is convex. The latter inequality is assumed to hold.

So, at least for π_L close to $\widehat{\pi}_L$ the crowding-out condition holds.

For the case of $q_c < q_{LH}$, the incentives compatibility constraints for the Principal are:

$$\pi_H(q_{HH} - \alpha_H C(q_{HH})) \geq \pi_H(q_{LH} - \alpha_H C(q_{LH})) + (1 - \pi_H)(q_c - \alpha_H C(q_c))$$
$$\pi_L(q_{LH} - \alpha_L C(q_{LH})) + (1 - \pi_L)(q_c - \alpha_L C(q_c)) \geq \pi_L(q_{HH} - \alpha_L C(q_{HH}))$$

As in the previous case, the crowding-out condition will hold at least for π_L close to $\widetilde{q}_L$ if $\widehat{\pi}_L q_{HH} \geq q_c$. Substituting the expression for $\widehat{\pi}_L$ gives

$$\frac{q_c - \alpha_L C(q_c)}{[q_{HH} - \alpha_L C(q_{HH})] + [q_c - \alpha_L C(q_c)] - [q_{LH} - \alpha_L C(q_{LH})]} q_{HH} \geq q_c$$

which can be rearranged to

$$\alpha_L q_{HH} \left[\frac{C(q_{HH})}{q_{HH}} - \frac{C(q_c)}{q_c} \right] \geq (q_{LH} - q_c) \left[\alpha_L \frac{C(q_{LH}) - C(q_c)}{q_{LH} - q_c} - 1 \right] \tag{34}$$

The left-hand side term $\frac{C(q_{HH})}{q_{HH}} - \frac{C(q_c)}{q_c} > 0$ because $q_{HH} > q_{LH} > q_c$.

The right-hand side term $\frac{C(q_{LH}) - C(q_c)}{q_{LH} - q_c} < 1$, because it's a slope of the secant line to the graph of the increasing convex function $C(q)$, which is lower than the slope of the tangent line at the right edge of the interval $[q_c, q_{LH}]$, $C'(q_{LH})$, for which we have $C'(q_{LH}) = \alpha_L \beta_H < 1$.

So, the right-hand side in (34) is positive, the left-hand side is negative, and, consequently, the inequality (34) holds.

□

Proof of Proposition 3

Proof. The proposition is established by checking the equilibrium conditions case by case.

Consider the no-control pooling equilibrium candidate. Each of the Principals shouldn't have an incentive to deviate to control, in which case the output q_c will be obtained. For the pooling equilibrium the Agent's beliefs on the Principal's type on the equilibrium path is $E\alpha$, and the pro-social Agent will perform at the level q_{EH}, determined by $C'(q_{EH}) = E\alpha\beta_H$, the selfish Agent will perform at level $q = 0$. So, the two Best Response conditions for the two types of Principal are

$$V_H = \pi_H(q_{EH} - \alpha_H C(q_{EH})) \geq q_c - \alpha_H C(q_c)$$
$$V_L = \pi_L(q_{EH} - \alpha_L C(q_{EH})) \geq q_c - \alpha_L C(q_c)$$

The two inequalities hold for small q_c, since the right-hand sides are equal to 0 for $q_c = 0$. The first condition which becomes binding for small q_c determines the threshold q_1.

For the large q_c, the right-hand sides of the two inequalities are negative. By decreasing q_c, the inequality for V_L becomes bonding and determines the threshold q_6.

Consider the control pooling equilibrium. The Principals' Best Response conditions are (the Agent will reasonable believe that the Principal deviating to no-control should be the selfish one)

$$V_H = q_c - \alpha_H C(q_c) \geq \pi_H(q_{HH} - \alpha_H C(q_{HH}))$$
$$V_L = q_c - \alpha_L C(q_c) \geq \pi_L(q_{HH} - \alpha_L C(q_{HH}))$$

The inequality for the V_H is stronger and determines the lower and upper bounds for the values of $q_c \in [q_3, q_4]$ for which the control pooling equilibrium emerges.

The case of the separating equilibrium is partially considered in Proposition 2. The conditions for the separating equilibrium to the right of the control pooling region are the same, and the "right-hand side" separating equilibrium emerges due to non-monotonicity of the payoff functions. The regions for q_c are $[q_2, q_3]$ on the left and $[q_4, q_5]$ on the right.

So, all the possible pure strategies equilibria are considered. The regions of the values of q_c not covered by the pure strategies equilibria, should bring the mixed strategies equilibria.

□

Proof of Lemma 5

According to (25), for any Best Response acceptance rule $a(\cdot)$ holds

$$a(q) \in \underset{a\in[0,1]}{\arg\max}\, (B - C(q) + \widehat{\alpha}(q)\beta q)\, a$$

The solution to this program is easy to find. If, for some q, $B - C(q) + \widehat{\alpha}(q)\beta q > 0$ then $a(q) = 1$, if $B - C(q) + \widehat{\alpha}(q)\beta q < 0$, then $a(q) = 0$ for the corresponding values of q, finally, if $B - C(q) + \widehat{\alpha}(q)\beta q = 0$, then any a is a solution .

According to lemma 4, the participation constraint $B - C(q) + \widehat{\alpha}(q)\beta q > 0$ is equivalent to the threshold $q < q^0(\widehat{\alpha}(q), \beta)$. This gives the characterization of the Best Response rule.

Since $\mu(q) \in [0, 1]$, then $\alpha_L \le \widehat{\alpha}(q) \le \alpha_H$, consequently, according to monotonicity of the function $q^0(\alpha, \beta)$ with respect to its first argument (see lemma 4) we obtain

$$q^0(\alpha_L, \beta) \le q^0(\widehat{\alpha}(q), \beta) \le q^0(\alpha_H, \beta)$$

which can be rewritten as $q^0_L \le q^0(\widehat{\alpha}(q), \beta) \le q^0_H$. This proofs (27).

Proof of Lemma 6

Proof. Since q^*_j $(j = L, H)$ are the elements of the Best Response of the Principal, the two inequalities hold:

$$(q^*_L - \alpha_L C(q^*_L))a^*(q^*_L) \ge (q^*_H - \alpha_L C(q^*_H))a^*(q^*_H)$$
$$(q^*_H - \alpha_H C(q^*_H))a^*(q^*_H) \ge (q^*_L - \alpha_H C(q^*_L))a^*(q^*_L)$$

Summing them up gives

$$(\alpha_H - \alpha_L)(C(q^*_L)a^*(q^*_L) - C(q^*_H)a^*(q^*_H)) \ge 0$$

which proves the first statement in the Lemma.

For the second claim of the Lemma, notice that if $a^*(q^*_H) = 1$, then

$$C(q^*_L) \ge C(q^*_L)a^*(q^*_L) \ge C(q^*_H)$$

which gives $q^*_L \ge q^*_H$. $\square$

Proof of Lemma 7

Proof. Let $q', q'' \in \operatorname{supp} \sigma^*_H$ and $q', q'' \in \operatorname{supp} \sigma^*_L$ are the two offers made by both types in an equilibrium. Then, the two types should be indifferent between the two offers:

$$(q' - \alpha_H C(q'))a^*(q') = (q'' - \alpha_H C(q''))a^*(q'')$$
$$(q' - \alpha_L C(q'))a^*(q') = (q'' - \alpha_L C(q''))a^*(q'')$$

This gives

$$\alpha_H(C(q'')a^*(q'') - C(q')a^*(q')) = \alpha_L(C(q'')a^*(q'') - C(q')a^*(q')) = \\ = q''a^*(q'') - q'a^*(q') \quad (35)$$

The first equality gives $C(q'')a^*(q'') - C(q')a^*(q') = 0$. This, in turn, gives

$$\frac{a^*(q'')}{a^*(q')} = \frac{q'}{q''} = \frac{C(q')}{C(q'')}$$

But the second part of this equality can't hold for a convex function $C(q)$ if $q' \ne q''$. This finishes the proof. $\square$

Proof of Lemma 8

Proof. Consider the case $\beta > \beta_2$, then $q_H^P < q_L^0$ and, consequently the offer q_H^P has to be accepted with probability 1 on the equilibrium path. By definition, the offer q_H^P maximizes the α_H-type utility. This means that α_H-type will offer only $q = q_H^P$. Moreover, some offers, higher than q_H^P, are accepted with probability 1, and consequently, α_L-type will never be interested to pool at q_H^P. So, there can't be any equilibrium with a pooling part (i.e. pooling or semi-separating) for $\beta > \beta_2$.

Consider the case $\beta \leq \beta_2$ and the pooling equilibria.

According to Lemma 7, the pooling equilibrium offer q_p^* is unique.

First, prove the necessity of the three conditions:

$$q_p^* \leq q_E^0, \quad q_p^* \leq q_L^{00}, \quad q_p^* \geq q_L^0 \tag{36}$$

1. The beliefs should be consistent on the equilibrium path, which means that $\mu^*(q_p^*) = \Pi$. This gives $\widehat{\alpha}(q_p^*) = \Pi$ and the corresponding acceptance threshold $\widehat{q}(q_p^*) = q_E^0$, according to (22) and Lemma 5. So, to have the offer q_p^* accepted, it is necessary to have $q_p^* \leq q_E^0$.

2. If the inequality $q_p^* \leq q_L^{00}$ is violated, then α_H-type has a profitable deviation to $\tilde{q} = q_L^0 - \varepsilon$ for small enough ε. In fact, since the function $V(q, \alpha_H)$ is decreasing in q for $q > q_L^{00} > q_P^H$ and according to the definition of q_L^{00}, we have $V(q_p^*, \alpha_H) < V(q_L^{00}, \alpha_H) = V(q_L^0, \alpha_H)$. Consequently, since $V(q, \alpha_H)$ is increasing in q for $q < q_L^0 < q_H^P$, we have $V(q_p^*, \alpha_H) < V(q_L^0 - \varepsilon, \alpha_H)$ for small enough ε. Since $a^*(q_L^0 - \varepsilon) = 1$, we have then $V(q_p^*, \alpha_H)a^*(q_p^*) < V(q_L^0 - \varepsilon, \alpha_H)a^*(q_L^0 - \varepsilon)$.

3. If $q_p^* < q_L^0$, then there is a profitable deviation for both Principal's types to $q_p^* + \varepsilon$, since the latter offer is still accepted with probability 1.

The necessity of the condition $a^*(q_p^*) = 1$ for $q_p^* < q_E^0$ is evident.

Second, prove the sufficiency of (36).

Consider a profile $(q_p^*, a^*(\cdot); \mu^*(\cdot))$ with offer q_p^* satisfying (36), acceptance rule $a^*(q_p^*) = 1$, $a^*(q) = 0$ for $(q \neq q_p^*,\ q > q_L^0)$, supported by beliefs $\mu^*(q_p^*) = \Pi$, $\mu^*(q) = 0$ for $q \neq q_p^*$.

Then, any downward deviation for both Principal's type isn't profitable, since it's either rejected or is too large (below q_L^0). Any upward deviation is rejected, so isn't profitable, too. Accepting of q_p^* is the Best Response for the Agent, and beliefs are consistent.

Finally, the relation between q_L^0, q_L^P, q_E^0, determined by the thresholds β_j in Proposition 4, leads to the statement of the lemma for the pooling equilibria.

Consider the case $\beta \leq \beta_2$ and the semi-separating equilibria.

In any semi-separating equilibrium there is a unique (according to lemma 7) pooling offer q_p^*. Denote by $a_p^* = a^*(q_p^*)$ the probability of acceptance of this offer.

Assume that there exists an equilibrium offer $q_L^* \in \operatorname{supp} \sigma_L^* \setminus \sigma_H^*$ (not necessarily unique), made by α_L-type only. Let $a_L^* = a^*(q_L^*)$. Then, α_L-type should be independent between the offers q_p^* and q_L^*:

$$\left(q_p^* - \alpha_L C(q_p^*)\right) a_p^* = \left(q_L^* - \alpha_L C(q_L^*)\right) a_L^* \tag{37}$$

On top of this, since the offer q_p^* is made by α_H-type, according to the monotonicity lemma 6,

$$C(q_p^*) a_p^* \leq C(q_L^*) a_L^* \tag{38}$$

Taking (37) and (38) together, we obtain $\frac{C(q_L^*)}{q_L^*} \geq \frac{C(q_p^*)}{q_p^*}$ which, in turn, leads to $q_L^* \geq q_p^*$.

At the same time, the offer q_L^* lies in the separating part of the semi-separating equilibrium, so $\mu^*(q_L^*) = 0$ and to be accepted, there should be $q_L^* \leq q_L^0$. On the other hand, $q_p^* \geq q_L^0$, because otherwise there will be profitable deviation to $q_p^* + \varepsilon$ for both types. So, we have $q_L^0 \geq q_L^* \geq q_p^* \geq q_L^0$, which is impossible for $q_L^* \neq q_p^*$.

So, it's impossible to have an offer q_L^* in the separating part of the semi-separating equilibrium.

Consider now the possibility of having the separating part for α_H-type. Let $q_H^* \in \operatorname{supp} \sigma_H^* \setminus \operatorname{supp} \sigma_L^*$ be (one of) offers, made by the α_H-type only, and let $a_H^* = a^*(q_H^*)$.

As in the previous case, the indifference for α_H-type and the monotonicity condition from lemma 6 should hold:

$$(q_H^* - \alpha_H C(q_H^*))\, a_H^* = \left(q_p^* - \alpha_H C(q_p^*)\right) a_p^* \tag{39}$$

$$C(q_H^*) a_H^* \leq C(q_p^*) a_p^* \tag{40}$$

Taken together, these conditions lead to $\frac{C(q_H^*)}{q_H^*} \leq \frac{C(q_p^*)}{q_p^*}$ which means that $q_H^* \leq q_p^*$.

Clearly, both q_H^* and q_p^* can't be less than q_L^0, otherwise there will be a profitable deviation for both types to the offer, greater by ε, which has to be accepted with probability 1.

Since $q_p^* < q_E^0$ (to be accepted with non-zero probability), then $q_H^* < q_E^0 < q_H^0$ for all equilibrium offers q_H^*. This means that they are accepted with probability 1, $a^*(q_H^*) = 1$. Any equilibrium offer q_H^* should solve $\max_{q \in Q} V(q, \alpha_j)\, a^*(q)$, and, since $a^*(q_H^*) = 1$, they should solve $\max_{q \in Q} V(q, \alpha_j)$. However, a solution of this program can consist of at most two points because of inverted-U shape of $V(q)$. Moreover, for $\beta \leq \beta_3$ the function $V(q, \alpha_H)$ is an increasing function in $q \in [0, q_E^0]$, so the solution of the program is unique, and consequently, there can be only one equilibrium offer q_H^* in the separating part for α_H-type.

This finishes the proof for the semi-separating equilibrium.

Separating equilibria For the case $\beta > \beta_2$ the offer q_H^P is accepted with probability 1, since $q_H^P < q_L^0$, so α_H-type will offer q_H^P and get $V(q_H^P, \alpha_H) = \max_q V(q, \alpha_H)$. Consider now the optimal offer for α_L-type. For $\beta > \beta_1$ the offer q_L^P is accepted with probability 1, since $q_L^P < q_L^0$, so α_L-type will offer q_L^P and get $V(q_L^P, \alpha_L) = \max_q V(q, \alpha_L)$. For the case $\beta_2 < \beta \leq \beta_1$, α_L-type will make the maximal acceptable offer, which is q_L^0. It has to be accepted with probability 1 on the equilibrium path, because otherwise there will be profitable deviation for α_L-type to an offer $\widetilde{q} = q_L^0 - \varepsilon$ with small enough ε (in fact, the offer $\widetilde{q}$ has to be accepted with probability 1, since $\widetilde{q} < q)L^0$, and $V(\widetilde{q}, \alpha_L) > V(q_L^0, \alpha_L) a^*(q_L^0)$ if $a^*(q_L^0) < 1$).

When $\beta \leq \beta_2$, in a separating equilibrium there should be at least two different offers $q_L^* \in \operatorname{supp} \sigma_L^*$ and $q_H^* \in \operatorname{supp} \sigma_H^*$, accepted with probabilities a_L^* and a_H^* respectively. The incentive compatibilities constraints for them are

$$(q_H^* - \alpha_H C(q_H^*)) a_H^* \geq (q_L^* - \alpha_H C(q_L^*)) a_L^*$$
$$(q_L^* - \alpha_L C(q_L^*)) a_L^* \geq (q_H^* - \alpha_L C(q_H^*)) a_H^*$$

which gives

$$\frac{q_L^* - \alpha_L C(q_L^*)}{q_H^* - \alpha_L C(q_H^*)} \geq \frac{a_H^*}{a_L^*} \geq \frac{q_L^* - \alpha_H C(q_L^*)}{q_H^* - \alpha_H C(q_H^*)}$$

and, after rearrangements, $\frac{C(q_L^*)}{q_L^*} \geq \frac{C(q_H^*)}{q_H^*}$, which means that $q_L^* \geq q_H^*$. In a separating equilibrium there should be $q_L^* > q_H^*$ for any equilibrium offers.

Any equilibrium offer q_L^* needs to be accepted with non-zero probability, which means that $q_L^* \leq q_L^0$. All the offers $q_L^* < q_L^0$ are accepted with probability 1. If $q_L^* = q_L^0$ and $a^*(q_L^0) < 1$, then there is a profitable deviation for α_L-type to $q_L^0 - \varepsilon$. So, any equilibrium offer q_L^* should be accepted with probability 1.

However, since all the equilibrium offers q_H^* are below all the offers q_L^* and the function $V(q, \alpha_H)$ is increasing for $q \in [q, q_L^0]$, then there is a profitable deviation for α_H-type to one of the offers $q_L^* \in \operatorname{supp} \sigma_L^*$.

This means that there can't be any separating equilibrium for $\beta \leq \beta_2$. □

Proof of Proposition 5

Proof. Since $q_L^0 \leq q_L^P$, both functions $V(q, \alpha_H)$ and $V(q, \alpha_L)$ are increasing in $q \in [0, q_L^0]$.

Consider pooling equilibria. According to Lemma 8, $q_p^* \leq q_E^0$ holds for any pooling PBE.

Check the intuitive criterion. Let q be a deviation such that $q < q_E^0$. Then

$$\max_{a \in BR(\mathcal{A}, q)} V(q, a, \alpha_j) = V(q, \alpha_j) \cdot 1$$

because, at least for the beliefs $\mu(q) = \Pi$ the offer q should be accepted for any reasonable acceptance rule - see Lemma 5.

Notice also that

$$\min_{a \in BR(\mathcal{A}', q)} V(q, a, \alpha_j) = 0 \text{ for } \mathcal{A}' = \{\alpha_L\} \text{ or } \mathcal{A}' = \mathcal{A}$$
$$\min_{a \in BR(\mathcal{A}', q)} V(q, a, \alpha_j) = V(q, \alpha_j) \text{ for } \mathcal{A}' = \{\alpha_H\}$$

because if beliefs are concentrated on the subset which contains α_L, then beliefs $\mu(q) = 0$ are possible and any offer $q > q_L^0$ (which is the case here) is rejected for any reasonable acceptance rule, so the minimal value of V is zero. On the other hand, if beliefs are concentrated on α_H, i.e. $\mu(q) = 1$, the offer $q < q_H^0$ (which is the case since $q < q_E^0 < q_H^0$) is accepted.

So, the only possibility to have the intuitive criterion (29) violated is to have $\mathcal{A} \setminus J(q) = \{\alpha_H\}$ and $V_H^* < V(q, \alpha_H)$. In other words, α_H-type should be "reasonably" revealed by deviation to q (see the definition of the set $J(q)$) and this deviation should be profitable, i.e. q should be closer to q_H^P, compared to the distance between q_p^* and q_H^P.

Three cases are possible as illustrated by figure 11: 1) $q_p^* \leq q_H^P$; 2) $q_H^P < q_p^* \leq q_L^P$; 3) $q_p^* > q_L^P$.

For the case 1 the deviations to $q < q_p^*$ are not profitable; the deviations to $q > q_p^*$ are profitable (at least for $q < q_L^P$) but they are profitable for both types or only for α_L-type. The deviations to $q > q_L^P$ can be profitable for α_L-type only. Consequently, any deviation to $q > q_p^*$ can't reasonably reveal α_H-type. So, all the equilibrium from area 1 pass the intuitive criterion.

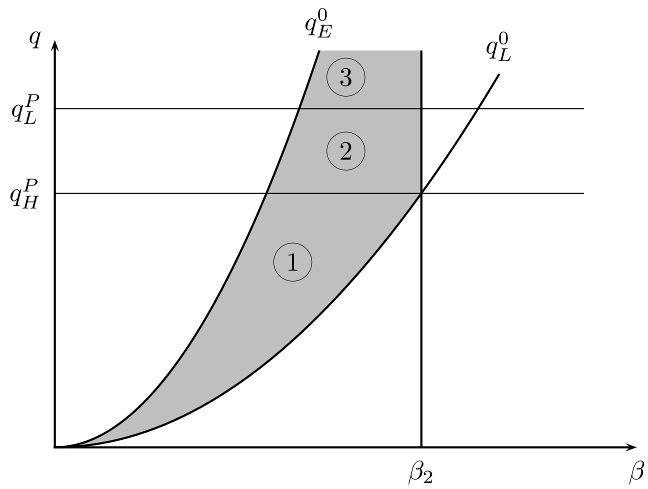

Figure 11:

For the case 2 the deviation to $q = q_H^P$ is profitable for α_H-type and isn't profitable for the α_L-type, so α_H-type is revealed by such deviation and the equilibria from area 2 don't satisfy the intuitive criterion.

Finally, for the case 3, the revealing profitable deviation for the α_H-type is constructed in the following way. Since $q_p^* > q_L^P$ there exists $q_L' < q_L^P$ (and then $q_L' < q_p^*$) such that $V(q_L', \alpha_L) = V(q_p^*, \alpha_L)$. Notice that q_L' is on the increasing part of $V(q, \alpha_L)$, so any $q < q_L'$ is not a profitable deviation for α_L. However, the deviation to q_L' is profitable for α_H-type (the proof is below), and then the deviation to $q = q_L' - \varepsilon$ is the revealing profitable deviation for α_H type.

Now I proof that the deviation to q_L' is profitable for α_H-type. Notice that the offers $q = q_L', q_p^*$ are accepted for any Best Response acceptance rule. So,

$$V(q, a, \alpha_j) = V(q, \alpha_j) \cdot 1 = q - \alpha_j C(q) + \alpha_j B$$

Since $V(q_L', \alpha_L) = V(q_p^*, \alpha_L)$, we have

$$q_p^* - \alpha_L C(q_p^*) = q_L' - \alpha_L C(q_L')$$

By using this, we get

$$\begin{aligned} q_L' - \alpha_H C(q_L') = q_L' - \alpha_L C(q_L') + (\alpha_H - \alpha_L) C(q_L') = \\ = q_p^* - \alpha_L C(q_p^*) + (\alpha_H - \alpha_L) C(q_L') = \\ = q_p^* - \alpha_H C(q_p^*) + (\alpha_H - \alpha_L) C(q_p^*) + (\alpha_H - \alpha_L) C(q_L') = \\ = q_p^* - \alpha_H C(q_p^*) + (\alpha_H - \alpha_L)(C(q_p^*) - C(q_L')) \end{aligned}$$

The second term is positive, so

$$q_L' - \alpha_H C(q_L') > q_p^* - \alpha_H C(q_p^*)$$

which gives the required inequality:

$$V(q_L', \alpha_H) > V(q_p^*, \alpha_H)$$

This finishes the analysis of case 3 and the pooling equilibrium case. We've got that only the equilibria from area 1 satisfy the intuitive criterion.

If an equilibrium is semi-separating, it can have one of the two structures: $q_H^* < q_p^*$ or $q_H^* < q_H^{**} < q_p^*$. Consider a deviation of the separating part of the equilibrium to q, which is closer to q_H^P: $q = q_H^* + \varepsilon$ if $q_H^* < q_H^P$, $q = q_H^* - \varepsilon$ if $q_H^* > q_H^P$ (it's impossible to have $q_H^* = q_H^P$ because then α_H-type strongly prefers q_H^P to the pooling part of the equilibrium candidate q_p^*, which can't be the case in equilibrium). First, $J(q) = \{\alpha_L\}$ because α_L-type is concentrated on the pooling part of the equilibrium q_p^*; since $q < q_p^*$ and $V(q, \alpha_L)$ is increasing, the deviation to q is unprofitable for α_L-type. Second, in the intuitive criterion (29) $a \in BR(\{\alpha_H\}, q)$, which means that the offer q is accepted with probability 1: $a(q) = 1$ and then $V_H^* = V(q_H^*, \alpha_H) < \min\limits_{a \in BR(\{\alpha_H\}, q)} V(q, a, \alpha_H)$ with $V(q, a, \alpha_H) \equiv V(q, \alpha_H)$).

So, the semiseparating equilibria fails to satisfy the intuitive criterion because of deviation to q, described above for α_H-type. $\square$

Proof of Proposition 6

Proof. For the case 1 all the Principals have $\alpha > \alpha^\times$, then, according to Claim 4, the preferred output for all types is feasible. So, accepting all offers $q^P(\alpha)$ is reasonable Best Response for the Agent and gives to all types of the Principal their unconstrained maximal utility.

For the case 2, first proof that there exists unique solution to (31).

For $\alpha = \alpha^\times$ we have $E_{\alpha^\times}[\alpha] < \alpha^\times$, and, consequently $q^0\left(E_{\alpha^\times}[\alpha]\right) < q^0\left(\alpha^\times\right) = q^P(\alpha^\times) < q^P\left(E_{\alpha^\times}[\alpha]\right)$, so the left-hand side of (31) is smaller than than the right-hand side.

For $\alpha = \alpha_2$ holds $E_{\alpha^\times}[\alpha] = E\alpha$. So, (30) means that the left-hand side of (31) is more than the right-hand side.

Since both sides of (31) are continuous, there exists unique solution to this equation.

Second, consider the Agent's acceptance rule given the Principal's offer. Clearly, all offers $q > \widetilde{q}^0$ should be reasonable rejected, offers $q < \widetilde{q}^0$ should be reasonably accepted with probability 1. The offer $q = \widetilde{q}^0$ should be accepted with probability 1 on the equilibrium path because accepting it with probability $a < 1$ will make Principal's deviation to $\widetilde{q}^0 - \varepsilon$ profitable.

Finally, ant type of the Principal can't do better since those with $\alpha > \widetilde{\alpha}$ implement their preferred output, and those with $\alpha < \widetilde{\alpha}$ could do better only by implementing $q > \widetilde{q}^0$ which are rejected.

Case 3 is considered in the same way as the pooling part ($\alpha < \widetilde{\alpha}$) in case 2. $\square$

References

Klaus Abbink, Bernd Irlenbusch, and Elke Renner. The moonlighting game An experimental study on reciprocity and retribution. *Journal of Economic Behavior and Organization*, 42(2):265–277, 2000.

James Andreoni. Philanthropy, Handbook of Giving, Reciprocity and Altruism, SC. Kolm and J. Mercier Ythier, eds, 2006.

Charles Bellemare and Bruce Shearer. Gift Exchange within a Firm: Evidence from a Field Experiment. *Cahier de recherche/Working Paper CIRPÉE*, 7: 08, 2007.

Roland Bénabou and Jean Tirole. Intrinsic and Extrinsic Motivation. *Review of Economic Studies*, 70(3):489–520, 2003.

Roland Bénabou and Jean Tirole. Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678, 2006a.

Roland Bénabou and Jean Tirole. Belief in a Just World and Redistributive Politics. *Quarterly Journal of Economics*, 121(2):699–746, 2006b.

Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10(1):122–142, 1995.

Sandra H. Berry and David E. Kanouse. Physician Response to a Mailed Survey an Experiment in Timing of Payment. *Public Opinion Quarterly*, 51(1):102–114, 1987.

Gary E. Bolton and Axel Ockenfels. Does Laboratory Trading Mirror Behavior in Real World Markets. *Fair Bargaining and Competitive Bidding on EBay, March*, 2008.

Patrick Bolton and Mathias Dewatripont. *Contract Theory*. MIT Press, 2005.

Gary Charness, Ernan Haruvy, and Doron Sonsino. Social distance and reciprocity: An Internet experiment. *Journal of Economic Behavior and Organization*, 63(1):88–103, 2007.

In-Koo Cho and David M. Kreps. Signaling Games and Stable Equilibria. *Quarterly Journal of Economics*, 102(2):179–221, 1987.

Martin Dufwenberg and Uri Gneezy. Measuring Beliefs in an Experimental Lost Wallet Game. *Games and Economic Behavior*, 30(2):163–182, 2000.

Robert Dur. Gift Exchange in the Workplace: Money or Attention? *Tinbergen Institute Discussion paper*, 2008.

Tore Ellingsen and Magnus Johannesson. Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3):990–1008, June 2008.

Armin Falk. Gift Exchange in the Field. *Econometrica*, 75(5):1501–1511, 2007.

Armin Falk and Urs Fischbacher. A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315, 2006.

Armin Falk and Michael Kosfeld. The Hidden Costs of Control. *American Economic Review*, 96(5):1611–1630, 2006.

Ernst Fehr and Armin Falk. Wage Rigidity in a Competitive Incomplete Contract Market. *Journal of Political Economy*, 107(1):106–134, 1999.

Ernst Fehr and Urs Fischbacher. Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2):63–87, 2004.

Ernst Fehr and John A. List. The Hidden Costs and Returns of Incentives-Trust and Trustworthiness Among CEOs. *Journal of the European Economic Association*, 2(5):743–771, 2004.

Ernst Fehr and Bettina Rockenbach. Detrimental effects of sanctions on human altruism. *Nature*, 422:137–140, 2003.

Ernst Fehr, George Kirchsteiger, and Arno Riedl. Does Fairness Prevent Market Clearing? An Experimental Investigation. *Quarterly Journal of Economics*, 108(2):437–460, 1993.

Ernst Fehr, Simon Gächter, and George Kirchsteiger. Reciprocity as a Contract Enforcement Device: Experimental Evidence. *Econometrica*, 65(4):833–860, 1997.

Drew Fudenberg and Jean Tirole. *Game Theory*. Mit Press, 1991.

Simon Gächter and Armin Falk. Reputation and Reciprocity: Consequences for the Labour Relation. *Scandinavian Journal of Economics*, 104(1):1–26, 2002.

Uri Gneezy. Does high wage lead to high profits? An experimental study of reciprocity using real effort. *Graduate School of Business, University of Chicago*, 2002.

Uri Gneezy and John A. List. Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments. *Econometrica*, 74(5):1365–1384, 2006.

Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3(4):367–88, 1982.

Heike Hennig-Schmidt, Bettina Rockenbach, and Abdolkarim Sadrieh. In search of workers real effort reciprocity–A field and a laboratory experiment. *Governance and the Efficiency of economic SYstems DP*, 55, 2005.

Daniel Kahneman, Jack L. Knetsch, and Richard Thaler. Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *American Economic Review*, 76(4):728–741, 1986.

Sebastian Kube, Michel André Maréchal, and Clemens Puppe. Putting Reciprocity to Work - Positive Versus Negative Responses in the Field. *SSRN eLibrary*, 2006.

David K. Levine. Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics*, 1(3):593–622, 1998.

John A. List. Field Experiments: A Bridge Between Lab and Naturally-Occurring Data. *NBER Working Paper*, 2007.

John A. List and Steven D. Levitt. What do laboratory experiments tell us about the real world. *NBER Working Paper*, 2005.

George Loewenstein, Ted O'Donoghue, and Matthew Rabin. Projection Bias In Predicting Future Utility. *Quarterly Journal of Economics*, 118(4):1209–1248, 2003.

Michel A. Maréchal and Christian Thöni. Do Managers Reciprocate? Field Experimental Evidence from a Competitive Market. *SSRN eLibrary*, 2007.

Harry J. Paarsch and Bruce S. Shearer. The Response to Incentives and Contractual Efficiency: Evidence from a Field Experiment. *Cahier de recherche CIRPE/Working Paper*, 7:01, 2007.

Canice Prendergast. The Provision of Incentives in Firms. *Journal of Economic Literature*, 37(1):7–63, 1999.

Matthew Rabin. Incorporating Fairness into Game Theory and Economics. *American Economic Review*, 83(5):1281–1302, 1993.

Bruce Shearer. Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment. *Review of Economic Studies*, 71(2):513–534, 2004.

Dirk Sliwka. Trust as a signal of a social norm and the hidden costs of incentive schemes. *American Economic Review*, 97(3):999–1012, 2007.

Jean Tirole. Rational irrationality: Some economics of self-management. *European Economic Review*, 46(4-5):633–655, 2002.