UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

WORKING PAPERS IN ECONOMICS

No 421

# Contracting Under Reciprocal Altruism

Oleg Shchetinin

December 2009

ISSN 1403-2473 (print)
ISSN 1403-2465 (online)

Department of Economics
School of Business, Economics and Law at University of Gothenburg
Vasagatan 1, PO Box 640, SE 405 30 Göteborg, Sweden
+46 31 786 0000, +46 31 786 1326 (fax)
www.handels.gu.se    info@handels.gu.se

# Contracting under Reciprocal Altruism *

Oleg Shchetinin†

December 4, 2009

**Abstract**

I show that a simple formal model of reciprocal altruism is able to predict human behavior in contracting situations, puzzling when considered within selfishness assumption. For instance, motivation and performance crowding-out are explained by a signaling mechanism in which provision of an extrinsic incentive signals non-generosity of the Principal and decreases Agent's intrinsic motivation. The model's equilibrium predicts behavior in the Control Game of Falk and Kosfeld and in a variant of Trust Game by Fehr and Rockenbach. This suggests that reciprocal altruism modeling could be fruitful more generally in applications of contract theory.

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

**JEL Classification Numbers:** D82, M54

*This paper is based to large degree on a part of my Ph.D. thesis, completed in Toulouse School of Economics. I'm grateful to Jean Tirole for many insightful discussions. I'm also grateful to Stefanie Brillon, Andrey Bremzen, Martin Dufwenberg, Robert Dur, Georg Kirchsteiger, Arjan Non, Dirk Sliwka, the participants of the BEE workshop in Toulouse, seminars at ECARES(Brussels), Gothenburg University, CEFIR(Moscow), participants of the 24-th Congress of the European Economic Association in Barcelona, the European Meeting of the Economic Science Association in Innsbruck and the 4-th Nordic Conference on Behavioral and Experimental Economics in Oslo for their comments. All errors are mine.

†Gothenburg University, School of Business, Economics and Law, Department of Economics.

# 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. Bénabou and Tirole (2003) and Bénabou and Tirole (2006b) argue that intrinsic motivation is important, whereas the provision of extrinsic incentives affects intrinsic motivation and shapes behavior in many different contexts.

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[1], 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. More precisely, a person is more altruistic towards 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

[1] See Sobel (2005) and Fehr and Schmidt (n.d.) for survey of theoretical literature on reciprocity and altruism. 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). 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.

game in which the Principal signals his altruism through offering a "generous" contract.

Broadly taken, my study contributes to behavioral theory of incentives. The model of reciprocal altruism is extended to encompass extrinsic incentives[2]; in this paper I focus on control and punishment for bad performance. I analyze the interaction between extrinsic incentives and intrinsic motivation[3], which can lead to the motivation crowding-out, explained here by the signaling mechanism. I show that the equilibrium structure of the emerging signaling games depends on the power of the available extrinsic incentive and obtain the conditions for crowding-out to emerge in equilibrium.

The following two assumptions are important in my analysis. First, the population of the Principals and the Agents is assumed to be heterogenous: together with selfish actors there are pro-social ones, who are more altruistic and reciprocal. The share of the pro-social actors is not known, but the actors have some beliefs about the population composition. Second, I assume that the actors believe that the rest of the population is "like themselves", i.e. they exhibit rational projection bias, "tendency to look at others...from the point of view of one's current self" (see Tirole (2002)).

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

[2] 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.

[3] 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 (2006b). 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.

The first variant follows the Trust Game of Fehr and Rockenbach (2003). The Principal chooses whether to punish the Agent for low performance, providing in this way extrinsic incentive, and sets the high-performance cutoff. After receiving the contract, the Agent can accept or reject it. In the experiment the Principals often choose not to punish, and, responding to this, many Agents choose to perform at a very high level. 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 for the selfish players case.

In the second variant, the Principal can either control the Agent by imposing lower bound for effort or give him full flexibility, so that zero effort is feasible. Such contract resembles the Control Game of Falk and Kosfeld (2006). In the experiment the Principals often choose not to control and, after this, many Agents perform at a very high level. If considered within the selfishness framework, the Principal's decision is a deviation from the equilibrium path, and many Agents deviate from the continuation subgame optimal move (zero effort).

I show that these "deviations" fit the equilibrium path of the proposed reciprocal altruism model. Intuitively, the reciprocal (pro-social) Agent's intrinsic motivation is boosted, and he performs at a high level when he's learned that the Principal is pro-social, or generous. The Principal can signal her generosity through offering a generous contract, i.e. not restricting the Agent or not threatening with punishment. However, the selfish Agent's intrinsic motivation can't be boosted, and he perform at the lowest possible level if not provided with extrinsic incentive. So, the observed performance of the Agents, not provided with extrinsic incentives, is either high or zero. On the other hand, the selfish Principals prefer to provide extrinsic incentive, revealing their types, and guarantee a relatively low performance from all Agents.

I argue that the model, despite its simplicity, accounts for the observed behavior surprisingly well, which justifies its relevance. This also suggests that the reciprocal altruism modeling can be fruitful more generally in applications of contract theory.

My model follows the general approach of Levine (1998) with pro-social component in utility depending on beliefs about partner's altruism,[4] which gives rise to signaling.

In a closely related paper Ellingsen and Johannesson (2008) propose a model, based on the taste for social esteem (pride) and unconditional altruism, incorporated in the utility function, leading to reciprocal behavior. In my model reciprocity is modeled in a more direct way, assuming that the Agent is a conditional altruist. Ellingsen and Johannesson (2008) propose a mechanism of crowding-out, different from mine.

Sliwka (2007) develops a model explaining reciprocal behavior, based on social norms. Together with unconditionally selfish and pro-social agents, there are conformists, whose utility depends on their beliefs about the shares of selfish and pro-social agents in the population. By proposing a generous contract, the Principal signals his conviction that the pro-sociality is relatively common, and, as a consequence, conformists turn to pro-sociality. In my model a generous contract signals the Principal's generosity, whereas in the model of Sliwka it provides information about the composition of the population.

My model is, however, simpler, compared to these two models and, probably, easier to extend. I also give a more structured description of the set of parameters under which the crowding-out equilibrium emerges.

The literature proposes a few theories of reciprocity, based on psychological games (see Rabin (1993), Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006)). In these models utility of a player depends not only on his material payoffs, but also on the perceived intentions of another player. In these models reciprocity is endogeneized, whereas I just assume that there are reciprocal agents. While the models, based on psychological games can be applied to explain behavior in the experiments, the analysis is complex and their direct application to contracting situations can be complicated. This literature, however, justifies incorporating reciprocity in a direct

[4] See also Dur (2008) for a model of reciprocal altruism, similar to mine.

way into preferences.

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 studies in detail application of the reciprocal altruism model to the experimental setting of Fehr and Rockenbach (2003) and Falk and Kosfeld (2006). I finish with some further thoughts in section 4.

## 2 The Reciprocal Altruism Framework

Consider a Principal-Agent relationship. The Principal is altruistic towards the Agent and the Agent reciprocates her altruism: if the Agent perceives the Principal to care about him, he becomes more altruistic towards her. 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 the 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

[5] More generally, $B$ can be treated as an opportunity cost of interacting with the Principal, not 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.

of the Principal and the Agent are then given by

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

Let $\alpha$ be the degree of the Principal's altruism and $\widehat{\alpha}$ denote the Agent's perception of the Principal's altruism. Let $\beta$ 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 $\beta\widehat{\alpha}$ represents the Agent's altruism emerging as a result of reciprocating altruism of the Principal[7].

Assume that $\alpha \in [\alpha_1, \alpha_2] \subseteq [0, 1]$ and $\beta\alpha_2 \leq 1$. The assumptions guarantee 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 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

[7] More generally, one can consider Agent's altruism of the form $\gamma(\widehat{\alpha})$ where $\gamma$ is an increasing function.

as functions of output are depicted in Figure 1.

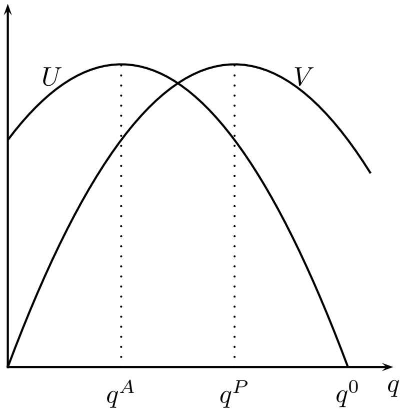

Figure 1: Principal's and Agent's payoffs under reciprocal altruism.

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 $\alpha$ or $\beta$, i.e. weaker internalization, there is a conflict of interest like in the standard Principal-Agent setup but this conflict is softened 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.

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 $\alpha$ 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 $\alpha$ and $\beta$ and on the information structure.

## 2.1 Benchmark cases

The preferred output for the Principal is given by

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

leading to

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

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 $\alpha$: $\frac{\partial q^P}{\partial \alpha} < 0$.*

The Lemma follows directly from (3)

The preferred value of output for the Agent is given by

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

leading to

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

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 the level such that marginal cost is equal to 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 $q^A(\widehat{\alpha}\beta)$ is determined by* (4) *and is an increasing function.*

The lemma follows directly from (4).

For the case of $\alpha < 1$ 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 (notice that $\alpha = \widehat{\alpha}$ is not required). This gap is larger, the smaller $\alpha$, $\widehat{\alpha}$ and $\beta$.

**Lemma 3.** *The Principal's preferred output is always larger than the Agent's one, except when it is known that $\alpha = \beta = 1$, in which case the preferred outputs are the same: $q^P(\alpha) > q^A(\widehat{\alpha}\beta)$, unless $\alpha = \widehat{\alpha} = 1$ and $\beta = 1$; $q^P(1) = q^A(1)$.*

*If $\alpha$ is known, $\widehat{\alpha} = \alpha$, then $\frac{\partial\left(q^P - q^A\right)}{\partial\alpha} < 0$, $\frac{\partial\left(q^P - q^A\right)}{\partial\beta} < 0$.*

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

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

**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

## 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, which 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 in this case $\widehat{q}$ is a "cheap talk" .

[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.

In the *Incentive treatment* the Principal, on top of $x$ and $\widehat{q}$, announces a fine $f$, imposed on the Agent if the back-transfer is lower than the desired level $\widehat{q}$, and so $\widehat{q}$ is no more a "cheap talk". The fine isn't paid to the Principal, it simply reduces the Agent's payoff, i.e. the fine is simply 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 study finds that, on average, the back-payment is higher when the Principal chooses not to punish ($f = 0$) than for the case of punishing ($f = \overline{f}$), in other words, providing an extrinsic incentive leads to a lower performance.

I show that the observed crowding-out in performance is an equilibrium outcome in the game, when utilities are determined, following the reciprocal altruism framework. In the considered experimental setting[9]

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

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 Agent decides on $q$. We can consider $x$ as a constant at this point and simplify the expressions for utilities of the players:

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

[9] In the experiment the monetary cost of paying back is linear: $C_m(q) = q$. One can assume that Principal's utility from money is concave with linear cost. Then, after rescaling utility to linear, cost become convex.
Alternatively, it can be assumed that there is also a psychological cost of paying back $C_\psi(q)$ which is 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.

[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.

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 $\theta^P$, and the type of the Agent by $\theta^A$. For both - the Principals and the Agents, $\theta^j \in \{\text{Social, Selfish}\}$. The type is private information.

The pro-social actors are characterized by altruism $\alpha_H$ and reciprocity intensity $\beta_H$, the selfish ones - by the pair $(\alpha_L, \beta_L)$, 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 $\beta_L = 0$.

In the considered game 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, whereas the behavior of the Agent is driven by his reciprocity, which is reflected by the structure of altruism in the utility functions $V$ and $U$ in (6) and (7). This setting can be generalized[12], but I stick to the simplest setting, capturing the idea of reciprocal altruism.

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 believe that the others in the society (or population) are like themselves, i.e. they exhibit rational projection bias. Loewenstein et al. (2003) provide evidence for the existence of the projection bias and develop a formal model. Bénabou and Tirole (2006a) discuss the implication of the projection bias for collective beliefs.

[11] A more general setting with the four possible pairs $(\alpha_k, \beta_l)$ can be considered. This, however, doesn't bring additional intuition. So, I restrict attention to a simpler setting.

[12] 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 $\alpha_p$ 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. Similarly, the Agent's altruism can be assumed to be equal to $\alpha_j + \beta_j \widehat{\alpha}$ with $j = L, H$, resulting in more than 2 values after observing the Principal's move in the case of separation.

Denote by $\pi_H$ the probability, assigned by the pro-social Principal to being matched with the pro-social Agent and by $\pi_L$ the probability, assigned to the same event by the selfish Principal:

$$\pi_H = Prob(\theta^A = \text{Social}|\theta^P = \text{Social}) = Prob(\beta = \beta_H|\alpha = \alpha_H) \qquad (8)$$

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

The projection bias assumption means that $\pi_L < \pi < \pi_H$ where $\pi$ is the true share of the pro-social actors.

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

**Game (T)**

The Principal is of type $i = H(L)$, i.e. $\theta^P = Social(Selfish)$, or, equivalently, $\alpha = \alpha_H(\alpha_L)$. The Agent is of type $j = H(L)$, i.e. $\theta^A = Social(Selfish)$, or, equivalently, $\beta = \beta_H(\beta_L)$. The types are privately known.

The Principal's strategy is a type-contingent pair $(f_i, \widehat{q}_i) \in \{0, \overline{f}\} \times [0, +\infty)$, $i = L, H$. 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 assigns probability $\pi_i$ to meeting the pro-social Agent. The Agent's ex-post beliefs $\mu$ 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 beliefs $\mu$ 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 $\mu$. The payoffs are given by (6) and (7).

The solution concept is Perfect Bayesian equilibrium[13], in which Agent's beliefs off the equilibrium path are "reasonable", in the sense of the intuitive criterion of Cho and Kreps[14].

Game (T) corresponds to the Incentive Treatment. For the Baseline

[13] A natural extension of the textbook version of PBE is needed (see, e.g. Fudenberg and Tirole (1991)), since we have incomplete information on both - Principal's and Agent's - sides.

[14] The refinement is needed only for the case of pooling equilibria, discussed in some detail in the analysis of the Control Game in subsection 3.2.

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. 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 $\widehat{\alpha}$, the Best Response back-transfer $q$ is:*

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

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

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

*where $\widetilde{q}^A(\widehat{\alpha}\beta)$ is an increasing function, determined by*

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

The proof of Claim 1 is given in the Appendix.

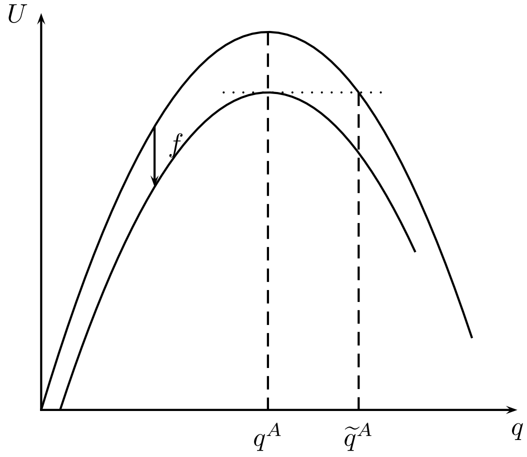

Figure 2: Agent's payoff

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, implementable with intrinsic motivation only (see Figure 2).

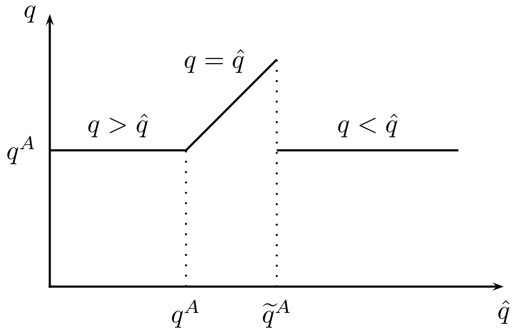

Figure 3: Back-transfer as a function of threshold

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".

It follows from the Claim that, contrary to the standard 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 for the intrinsically motivated Agent, so that he 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, for the case of $\widehat{q} > q^A$ the 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, implementable with intrinsic motivation only, given that the Agent with $\beta = \beta_j$ holds belief $\widehat{\alpha} = \alpha_i$. These back-transfers are determined by $C'(q_{ij}) = \alpha_i \beta_j$. The assumption $\beta_L = 0$ leads to $q_{HL} = q_{LL} = 0$.

Denote by

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

the maximal back-transfer, implementable when both intrinsic and extrinsic motivation are in place, 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

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

Finally, denote by $q^*_{ij}$ the equilibrium performance (back-transfer) for $\alpha_i$-Principal by $\beta_j$-Agent.

Call an equilibrium of game (T) to be separating equilibrium with crowding-out if (1) $\alpha_H$-type imposes no (threat of) fine, $\alpha_L$-type threatens with a fine: $f^*_H = 0$, $f^*_L = \overline{f}$ and (2) the average back-transfer to $\alpha_H$-type is higher than to $\alpha_L$-type: $\pi q^*_{HH} + (1-\pi) q^*_{HL} > \pi q^*_{LH} + (1-\pi) q^*_{LL}$.

In such equilibrium since generosity of the Principal is revealed and is then reciprocated by the pro-social Agent, he becomes intrinsically motivated to perform at the relatively high level $q_{HH}$. At the same time, the selfish Agent doesn't reciprocate, and, since the extrinsic incentive isn't provided, performs at zero level. On the other hand, the selfish Principal, by providing extrinsic incentive and signaling her low generosity (or toughness), can't intrinsically motivate the pro-social Agent at a high level, but guarantees (relatively low) performance $\widetilde{q}_{LL}$ from all the Agents[15]. The selfish Principal doesn't want to deviate to the unsure outcome "$q_{HH}$ or 0" because she is less confident in the possibility of inspiring high intrinsic motivation of the Agent, compared to the pro-social Principal.

I will focus on the case of $q_{LH} \leq \widetilde{q}_{LL}$ in the further analysis, as it simplifies the technical details. However, all the results, formally stated below, hold for the case of $q_{LH} \geq \widetilde{q}_{LL}$ as well; the required alterations are described in

[15] The pro-social Agent is intrinsically motivated to perform at $q_{LH}$ and will do so if $q_{LH} > \widetilde{q}_{LL}$. I will, however, assume that $q_{LH} \leq \widetilde{q}_{LL}$ to exclude such possibility, as it complicates the analysis without bringing new intuition.

the footnotes.

The incentives compatibility conditions are necessary for the existence of the separating equilibrium. They write as

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

The conditions are equivalent to the restrictions on the Principal's beliefs: $\pi_L \leq \widehat{\pi}_L$, $\pi_H \geq \widehat{\pi}_H$, with[16]

$$\widehat{\pi}_H = \frac{\widetilde{q}_{LL} - \alpha_H C(\widetilde{q}_{LL})}{q_{HH} - \alpha_H C(q_{HH})}, \qquad \widehat{\pi}_L = \frac{\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})}{q_{HH} - \alpha_L C(q_{HH})} \tag{11}$$

The selfish Principal can also consider deviation to requiring $\widehat{q} = \widetilde{q}_{LH}$ (notice that $\widetilde{q}_{LH} > \widetilde{q}_{LL}$). In this case only the pro-social Agent would perform at the required level, whereas the selfish Agent would choose $q = 0$ and pay fine. However, if $\widetilde{q}_{LH} > q_{HH}$, the selfish Principal would be better off by choosing $\widehat{q} = \widetilde{q}_{LH}$, compared to outcome in the separating equilibrium with crowding-out. So, the condition $\widetilde{q}_{LH} \leq q_{HH}$ is also necessary for the existence of the separating equilibrium with crowding-out.

It turns out that these conditions are sufficient for the existence of the separating equilibrium and for the crowding-out in performance. To guarantee that the obtained equilibrium is unique, a stronger restriction on $\pi_L$ is needed.

[16] One can check that for the case of $q_{LH} \leq \widetilde{q}_{LL}$ the incentive compatibility constraints write as

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

and the corresponding beliefs cut-offs are

$$\widehat{\pi}_H = \frac{\widetilde{q}_{LL} - \alpha_H C(\widetilde{q}_{LL})}{[q_{HH} - \alpha_H C(q_{HH})] - [q_{LH} - \alpha_H C(q_{LH})] + [\widetilde{q}_{LL} - \alpha_H C(\widetilde{q}_{LL})]}$$
$$\widehat{\pi}_L = \frac{\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})}{[q_{HH} - \alpha_L C(q_{HH})] - [q_{LH} - \alpha_L C(q_{LH})] + [\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})]}$$

**Proposition 1.** *Assume that $q_{LH} \leq \widetilde{q}_{LL}$. Game (T) has a separating equilibrium with crowding-out iff*

$$\widetilde{q}_{LH} \leq q_{HH}, \quad \pi_H \geq \widehat{\pi}_H, \quad \pi_L \leq \widehat{\pi}_L \tag{12}$$

*where $\widehat{\pi}_H < 1$, $\widehat{\pi}_L > 0$.*

*The performance in equilibrium is*

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

*This equilibrium is the unique equilibrium of game (T) if $\pi_L \leq \widehat{\widehat{\pi}}_L$, where $\widehat{\widehat{\pi}}_L = \frac{\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})}{q^{\times} - \alpha_L C(q^{\times})}$.*

The Proof of Proposition 1 is given in the Appendix[17].

Crowding-out of intrinsic motivation is explained by the signaling mechanism. The provision of the extrinsic incentives can offset the crowding effect on performance, but for some values of parameters, as described by (12), the crowding-out of intrinsic motivation has a stronger negative effect on performance than the positive effect of an extrinsic incentive.

The strength of the available extrinsic motivator $\overline{f}$ influences the structure of the equilibrium of the game. Notice that if the available extrinsic motivator is very weak, i.e. $\overline{f}$ is small, then the guaranteed performance under the tough contract $\widetilde{q}_{LL}$ is small either. Then, given that $\pi_L$ is large enough, the selfish Principal prefers to deviate from the tough contract, and, consequently, the separating equilibrium with crowding-out can't emerge. On the other hand, if the available extrinsic motivator is very strong, i.e. $\overline{f}$ is large, then the Agent's performance when extrinsic incentive is provided can be high even if intrinsic motivation is weak. Then, the pro-social Principal can prefer using extrinsic incentive instead of signaling her generosity through offering the generous contract and separating equilibrium can't emerge. To sum up, the separating equilibrium with crowding-out can emerge only when the

[17] The Proposition remains to hold for the case of $\widetilde{q}_{LH} \leq q_{HH}$. In this case only changes the equilibrium performance $q^*_{LH} = q_{LH}$ and the thresholds for beliefs - see the previous footnote.

strength of the available extrinsic incentive is of some middle value. The formal statement follows from Proposition 1.

**Corollary 1.** *For any triple* $(\alpha_L, \alpha_H, \beta_H)$*, satisfying*

$$C(q_{LH}) \leq C(q_{HH}) - C(q_{LH}) - \alpha_L \beta_H (q_{HH} - q_{LH}) \tag{13}$$

*there exists a non-empty set* $M$ *of the parameters* $(\pi_L, \pi_H, \overline{f})$ *such that the unique equilibrium of the game (T) is the separating equilibrium with crowding-out. For any* $(\pi_L, \pi_H, \overline{f}) \in M$ *holds* $f_1 \leq \overline{f} \leq f_2$.

The Proof of Corollary 1 is given in the Appendix.

The condition (13) is a direct consequence of limiting the analysis to the case of $q_{LH} \leq \widetilde{q}_{LL}$. The condition can be relaxed when equilibria for $q_{LH} \geq \widetilde{q}_{LL}$ are included in the analysis as well.

Consider now the baseline treatment of the experiment with no possibility of imposing the fine. According to the model, separation can't emerge in this case[18], so there will be pooling equilibrium, in which the back-payment from the pro-social Agent is $q_{EH} = q^A(E\alpha \cdot \beta_H)$ and the selfish Agent pays back 0. Consequently, the average back-payment is $\pi q_{EH} < \pi q_{HH}$. The model then predicts that the back-payment to the pro-social Principal (not imposing the fine) in the incentive treatment is higher than the back-payment in the baseline treatment, exactly as it's observed in the experiment. The model shows that it happens exactly because of lack of signaling opportunities in the baseline treatment, so that intrinsic motivation of the Agent can't be boosted.

Finally, a numerical exercise complements the analysis and shows that the model can reliably predict the observed behavioral patterns with reasonable values of parameters. As reported by Fehr and Rockenbach (2003), in the incentive treatment when the fine isn't imposed, 7 out of 15 Agents chose high back-transfers (higher than 15) and the rest chose low back-transfers[19],

[18] In the experiment, the choice of the investment amount serves as a signal of the degree of altruism of the Principal. However, we focus here on a simpler model and don't take this into account.

[19] These back-transfers were greater than zero, contrary to the model's prediction. This, however, can be attributed to the fact that $\beta_L > 0$.

resulting in the average back-transfer of 12.5. To fit these data, one can choose $q_{HH} = 22$ and $q_{HL} = 4.15$, leading to the average performance of $\frac{7}{15} \cdot 22 + \frac{8}{15} \cdot 4.15 = 12.5$. When the fine is imposed, 10 out of 30 Agents pay back 0 with the rest paying some positive amounts with the average performance of 6. The back-transfers of zero can be attributed to too high required performance of Principals, which can be a result of an out-of equilibrium play or Principals' overconfidence in Agents' intrinsic motivation. To fit the data, one can take $\widetilde{q}_{LL} = 9$, leading to average performance of $\frac{2}{3} \cdot 9 + \frac{1}{3} \cdot 0 = 6$. Take $\alpha_H = 0.9$, $\alpha_L = 0.4$, $\beta_H = 0.9$. Set $\widehat{f} = 4$, as in the experiment. Consider cost function of the form $C(q) = a(q+b)^2 + c$. We can now find the values of $a, b, c$, satisfying to three conditions: $q_{HH} = 22$, $\widetilde{q}_{LL} = 9$, $C(0) = 0$. This gives $a = 0.0104$, $b = 16.7766$, $c = -2.9396$. Other relevant parameters are $\widehat{\pi}_H = 0.51$, $\widehat{\pi}_L = 0.44$ and $\pi = \frac{7}{15} = 0.47$, as follows from the Agents' response to the Principals' offers of the contract without fine. It's important to check that $\widetilde{q}_{LH} = 20.03 < q_{HH}$ to guarantee the existence of the separating equilibrium.

## 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 (payment) from below. Output is assumed to be equal to effort.

Put formally, the Principal offers a contract $\underline{q}$ which can take two values - 0 or $q_c > 0$, with $q_c$ 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 $\underline{q} = q_c > 0$, exert, on average, less effort, than when offered $\underline{q} = 0$, which means that extrinsic incentive (control) has on average a negative impact on Agents' performance. However, the observed behavior of the Agents is heterogenous: positive, negative and neutral reaction to control were all observed. Finally, there is heterogeneity among Principals: many of them choose not to control.

I show that the reciprocal altruism framework accounts for these behaviors. As for the Trust Game, I build a model matching the experimental design, following the reciprocal altruism framework, and show that the equilibrium of the emerging game coincides with the observed behaviors.

The selfish utilities of the Principal and the Agent are given by[20] $v = q$ and $u = 120 - C(q)$ respectively.

Taking into account the reciprocal and altruistic components leads to the (social) utilities $V = q + \alpha(120 - C(q))$ and $U = 120 - C(q) + \widehat{\alpha}\beta q$.

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

$$V = q - \alpha C(q) \tag{14}$$

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

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

**Game (C)**

The Principal is of type $i = H(L)$ if $\alpha = \alpha_H(\alpha_L)$, the Agent is of type $j = H(L)$ if $\beta = \beta_H(\beta_L)$. 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 level of performance, conditional on the Principal's action $q_j(\underline{q}) \in [\underline{q}, +\infty)$, $j = L, H$.

The Principal assigns probability $\pi_i$ to meet the pro-social Agent - see (8)-(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 $\mu$ 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 $\mu$. The payoffs are given by (14) and (15).

As in the analysis of the Trust game, I look for the Perfect Bayesian

[20] 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.

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, i.e. the lower bound $\underline{q}$ of the feasible efforts set.

Denote by $q_{ij}$ the effort, voluntarily exerted by the $\beta_j$-Agent believing that the Principal's type is $\alpha_i$, i.e. $q_{ij} = q^A(\alpha_i\beta_j)$ with $C'(q_{ij}) = \alpha_i\beta_j$, according to (4).

Call an equilibrium of game (C) to be the separating equilibrium with crowding-out if (1) $\alpha_H$-type doesn't control, $\alpha_L$-type controls, i.e. $\underline{q}_H^* = 0$, $\underline{q}_L^* = q_c$ and (2) the average performance to $\alpha_H$-type is higher than to $\alpha_L$-type.

In such equilibrium the uncontrolled pro-social Agent is highly intrinsically motivated, since $\alpha_H$ type is revealed, and performs at relatively high level $q_{HH}$. The uncontrolled selfish Agent can't be intrinsically motivated and since the extrinsic incentive isn't provided, performs at zero level. When controlled, $\alpha_L$-type is revealed and $\beta_H$-type is intrinsically motivated to perform at the relatively low level $q_{LH}$, then his performance is $q = \max\{q_{LH}, q_c\}$. The performance of the controlled selfish Agent is $q_c$.

Denote

$$q^{\times} = \max\{q_{LH}, q_c\}$$

For the existence of the separating equilibrium, the incentives compatibility conditions (IC) should hold. They write as[21]

$$\pi_H(q_{HH} - \alpha_H C(q_{HH})) \geq \pi_H(q^\times - \alpha_H C(q^\times)) + (1-\pi_H)(q_c - \alpha_H C(q_c)) \tag{16}$$

$$\pi_L(q^\times - \alpha_L C(q^\times)) + (1-\pi_L)(q_c - \alpha_L C(q_c)) \geq \pi_L(q_{HH} - \alpha_L C(q_{HH})) \tag{17}$$

The IC conditions are equivalent to the restrictions on the Principal's beliefs $\pi_L \leq \widehat{\pi}_L$, $\pi_H \geq \widehat{\pi}_H$, where[22]

$$\widehat{\pi}_H = \frac{q_c - \alpha_H C(q_c)}{[q_{HH} - \alpha_H C(q_{HH})] + [q_c - \alpha_H C(q_c)] - [q^\times - \alpha_H C(q^\times)]} \tag{18}$$

$$\widehat{\pi}_L = \frac{q_c - \alpha_L C(q_c)}{[q_{HH} - \alpha_L C(q_{HH})] + [q_c - \alpha_L C(q_c)] - [q^\times - \alpha_L C(q^\times)]} \tag{19}$$

Notice that if $q_c \geq q_{HH}$ then even if separating equilibrium emerges, it's impossible to have crowding-out, because all the controlled Agents perform at level $q_c$, which is higher than performance of uncontrolled Agents.

It turns out that these conditions are not only necessary, but also sufficient for the existence of the separating equilibrium with crowding-out. The equilibrium is the unique pure strategy equilibrium; an additional restriction is required to rule out mixed strategies equilibria.

**Proposition 2.** *The separating equilibrium with crowding-out is the unique pure-strategies equilibrium of game (C) iff*

$$q_c < q_{HH}, \quad \pi_L \leq \widehat{\pi}_L, \quad \pi_H \geq \widehat{\pi}_H$$

*where $\widehat{\pi}_i$ are given by (18), (19) and $\widehat{\pi}_L > 0$, $\widehat{\pi}_H < 1$.*

*The performance in the equilibrium is*

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

[21] For the case $q_c \geq q_{LH}$ IC simplify to $\pi_H(q_{HH} - \alpha_H C(q_{HH})) \geq q_c - \alpha_H C(q_c)$, $q_c - \alpha_L C(q_c) \geq \pi_L(q_{HH} - \alpha_L C(q_{HH}))$.

[22] For $q_c \geq q_{LH}$ the beliefs cut-offs are $\widehat{\pi}_L = \frac{q_c - \alpha_L C(q_c)}{q_{HH} - \alpha_L C(q_{HH})}$, $\widehat{\pi}_H = \frac{q_c - \alpha_H C(q_c)}{q_{HH} - \alpha_H C(q_{HH})}$.

*The equilibrium is the unique equilibrium of game (C) if*

$$\pi_H > \frac{q_c - \alpha_H C(q_c)}{[q_{HH} - \alpha_H C(q_{HH})] + [q_c - \alpha_H C(q_c)] - [q_{EH} - \alpha_H C(q_{EH})]} \tag{20}$$

*where* $q_{EH} = q^A(E[\alpha] \cdot \beta_H)$.

The Proof of Proposition 2 is given in the Appendix.

The crowding-out is explained by the signaling mechanism. By choosing not to control, the pro-social Principal signals her kindness, inspiring 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, because he isn't reciprocal, and, once not controlled, exerts zero effort. The selfish Principal chooses to control and guarantees the (comparatively low) output $q_c$.

As in the Trust Game, the separating crowding-out equilibrium emerges when the available extrinsic incentive is neither too weak nor too strong. The argument for this is similar to the one for the Control game: imposing a weak extrinsic incentive can't compensate crowding-out of intrinsic motivation and then $\alpha_L$-type prefers pooling on no-control even if some of the Agents perform at zero level in this case. On the other hand, if the available extrinsic incentive is strong enough, imposing it increases performance sufficiently to compensate crowding-out of intrinsic motivation.

I describe now the equilibrium structure of game (C) for all values of $q_c$.

**Proposition 3.** *For any given* $(\alpha_L, \alpha_H, \beta_H, \pi_L, \pi_H)$*, there exist the thresholds* $q_i$*,* $q_i \leq q_j$ *for* $i < j$*, such that the unique pure strategy equilibrium of game (C) is:*

1. *No-control pooling for* $q_c \in [0, q_1]$*;*
2. *Separating equilibrium with crowding-out for* $q_c \in [q_2, q_3]$*;*
3. *Control pooling for* $q_c \in [q_3, q_4]$*;*
4. *Separating with no crowding-out in effort* $q_c \in [q_4, q_5]$*;*

*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.

I now discuss the relation between the experimental results and the predictions of the model. First, it was found in the experiment that for small $q_c$ (5 and 10) most of the Principals pool on no-control (74% and 71% respectively), whereas the Principals' choice for $q_c = 20$ resembles a separating equilibrium: 48% of Principals choose to control and 52% choose not to control, which is in line with proposition 3 if $10 < q_1$ and $q_2 < 20 < q_3$, which are not restrictive conditions.

Second, consider Agents' responds to Principals' choices. Although Agents' reaction[23] was highly heterogenous, it can be summarized in the following way[24]. For each treatment ($q_c = 5, 10, 20$) in case of trust there are Agents, performing at low level $q \leq 5$ - they can be viewed as selfish, and those, performing at $q \geq 5$, which can be thought to be pro-social. For the control case, I distinguish between agents, performing at level $q \leq q_c + 2$ (close to the minimal available performance) and those performing at $q > q_c + 2$. So, the Agents can be classified as follows:

| Performance under trust | Performance under control | |
|---|---|---|
| | $q \leq q_c + 2$ | $q > q_c + 2$ |
| $q \leq 5$ | Selfish | |
| $q > 5$ | Crowding-out | Keeping intrinsic motivation |

The number of Agents in each category is shown in the following table:

| | Treatment | | | | | |
|---|---|---|---|---|---|---|
| | $q_c = 5$ | | $q_c = 10$ | | $q_c = 20$ | |
| Performance under trust | Performance under control | | | | | |
| | $q \leq 7$ | $q > 7$ | $q \leq 12$ | $q > 12$ | $q \leq 22$ | $q > 22$ |
| $q \leq 5$ | 12 | 2 | 15 | 0 | 16 | 1 |
| $q > 5$ | 26 | 30 | 28 | 28 | 30 | 20 |

[23] In the experiment the strategy method was used, so for each Agent the performance for both control and trust was elicited.

[24] I use the detailed data from the Appendix of the paper.

Notice that selfish agents perform at the minimal possible level (or close to it) under control as well as under trust, with only few exceptions. Among the pro-social agents there are those, whose intrinsic motivation is crowded out when they are controlled, and they perform at the minimal possible level. There are, however, many pro-social agents, performing at a high level, even if they are controlled. The behavior of the latter is driven by mechanisms, different from reciprocity (e.g. unconditional altruism or fairness). The behavior of the former can, clearly, be explained by reciprocal altruism. The following observation supports the idea that not using a stronger extrinsic incentive signals higher altruism of the Principal, boosting intrinsic motivation of the Agent to a larger degree. The pro-social agents with crowded-out intrinsic motivation perform on average at $q = 24$ (median 22.5) when $q_c = 10$, and at $q = 33.7$ (median 33) when $q_c = 20$. The difference is statistically significant: Mann-Whitney z-statistic is $-2.354$, p-value is 0.186.

The observed behavioral choices (focus here on the case of $q_c = 20$) can be fitted by the model with cost function of the form $C(q) = a(q + b)^2 + c$, and $\alpha_H = 0.9$, $\alpha_H = 0.4$, $\beta_H = 0.9$ (these are arbitrary choices). The parameters of the cost function can be obtained by imposing the constraints $q_{HH} = 33$, $C(0) = 0$, $q_{LH} = 10$ (the last constraint is imposed only to guarantee that $q_{LH} < q_c$). The corresponding parameters values are $a = 0.00978$, $b = 8.4$, $c = -0.69$. Then the beliefs thresholds are $\widehat{\pi}_L = 0.64$, $\widehat{\pi}_H = 0.73$. The actual share of the pro-social Agents[25] is $\frac{30}{46} = 0.652$.

# 4 Concluding Remarks

In both applications of the reciprocal altruism framework, considered in this paper, the signaling mechanism and the existence of the separating equilibrium are crucial to explain the observed behavioral. Sorting conditions are crucial for the emergence of the separating equilibrium, so I discuss them now.

In the considered settings the Principal can offer two types of contracts -

[25] If the pro-social Agents with non-crowded intrinsic motivation are also taken into account, then the share is $\frac{50}{66} = 0.76$.

generous or restrictive, i.e. without imposing a painful extrinsic incentive or a contract comprising it. When the contract is generous ($f = 0$ in TG, or $\underline{q} = 0$ in CG), the Agent's performance is determined by his intrinsic motivation. In the restrictive contract, the Principal imposes extrinsic incentive, which restricts agent's choice of effort (direct restriction in CG, or punishment in case of low performance in TG). Agent's performance is the determined jointly by diminished intrinsic motivation and extrinsic incentive.

With the two-type setting, in which one Agent type is completely selfish, so that he can't be intrinsically motivated, only the pro-social Agent performs under the generous contract. Denote by $q_G$ his performance (it is equal to $q_{HH}$ in both TG and CG). Under restrictive contract, the selfish and pro-social agents can perform at different levels, delivering to the Principal expected utility $E_i[V_i(q_R)]$, where $i = L, H$ is the Principal's type.

The sorting condition writes then as

$$\pi_H V_H(q_G) - E_H[V_H(q_R)] \geq \pi_L V_L(q_G) - E_L[V_L(q_R)]$$

for all $q_R \leq q_G \leq Q$. The cut-off $Q$ is needed because of non-monotonicity of function $V$.

According to the sorting condition, if the selfish Principal prefers offering the generous contract, the pro-social Principal prefers to do so even stronger. Whereas, if the pro-social Principal prefers to offer the restrictive contract, the selfish Principal prefers to offer it to a larger degree. It's straightforward to check that the sorting conditions hold for the range of the parameters, sufficient for the existence of the separating equilibria in games (C) and (T).

Second, there is an important distinction between crowding-out in motivation and crowding-out in performance. For instance, crowding-out in motivation doesn't necessarily lead to crowding-out in performance. Hence, even if crowding-out in performance isn't observed, there can be crowding-out in motivation (workers can work hard, while being very unhappy and dissatisfied about how hard they work). This is the case for in Control Game for $q_c \in [q_4, q_5]$ (see Proposition 3). So, one should take care when relating lab or field evidence on performance to changes in motivation.

Finally, this paper shows that reciprocity is relevant in contracting situation as it influences intrinsic motivation of agents. The model of behavior, based on reciprocity, accounts for the observed behavioral patterns. This means, for instance, that taking into account reciprocity in theory of incentives can lead to new insights on the influence of (extrinsic) incentives on human behavior.

# 5 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 in both baseline treatment and incentive treatment when fine isn't imposed, and therefore chooses his preferred back-transfer.

For statement 2, notice that if $\widehat{q} \leq q^A$, then by choosing $q = q^A$ the Agent gets maximal utility and avoids paying the fine.

Consider the case of $\widehat{q} > q^A$. Notice that $\widetilde{q}^A(\widehat{\alpha})$ is constructed in such way that

$$\overset{\circ}{U}(q) > \overset{\circ}{U}(q^A) - f \quad \text{for} \quad q^A(\widehat{\alpha}, \beta) < \widehat{q} < \widetilde{q}^A(\widehat{\alpha}, \beta) \tag{21}$$

$$\overset{\circ}{U}(q) < \overset{\circ}{U}(q^A) - f \quad \text{for} \quad \widehat{q} > \widetilde{q}^A(\widehat{\alpha}, \beta) \tag{22}$$

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

It's clear that if (21) is the case, the Agent prefers to diverge from $q^A$ to $q > q^A$, whereas if (22) is the case, the Agent prefers to pay fine and

$q = \widehat{q}$. □

## Proof of the Proposition 1

*Proof.* **Existence.** Since Agent's beliefs are correct in the equilibrium, his performance is optimal, according to Claim 1. The separation is ensured by the incentive compatibility constraints (11) or (16).

To show the optimality of choosing $\widehat{q} = \widetilde{q}_{LL}$ for $\alpha_L$-type, notice, first, that $\widetilde{q}_{LL}$ is the maximal performance, implementable by imposing the fine and revealing $\alpha_L$-type when no agent perform at zero level (and pays fine). Second, if only $\beta_H$-Agent performs at the required level, then it's optimal for $\alpha_L$-type to set $\widehat{q} = \widetilde{q}_{LH}$, which leads to the expected utility $\pi_L V(\widetilde{q}_{LH}, \alpha_L)$. However, since it's assumed that $\widetilde{q}_{LH} \leq q_{HH}$, it follows that $V(\widetilde{q}_{LH}, \alpha_L) \leq V(q_{HH}, \alpha_L) \leq \pi_L V(\max\{q_{LH}, \widetilde{q}_{LL}\}, \alpha_L) + (1-\pi_L)V(q_{LL}, \alpha_L)$, where the latter inequality is the incentive compatibility constraint. This, however, means that it's better for $\alpha_L$-type to require performance $\widehat{q} = \widetilde{q}_{LL}$ and to guarantee that no agent prefers choosing $q = 0$.

**Inequality** $\widehat{\pi}_H \leq 1$ is equivalent to $\frac{C(q_{HH}) - C(\widetilde{q}_{LL})}{q_{HH} - \widetilde{q}_{LL}} \leq \frac{1}{\alpha_H}$, where the left-hand side $\frac{1}{\alpha_H} \geq 1$ and 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 $q = q_{HH}$, equal to $C'(q_{HH}) = \alpha_H \beta_H < 1$. So, $\widehat{\pi}_H \leq 1$ holds.

The inequality $\widehat{\pi}_L > 0$ is evident.

**Crowding-out condition** $\pi q_{HH} \geq \widetilde{q}_{LL}$**.** Since $\pi > \widehat{\pi}_L$, it is sufficient to prove that $\widehat{\pi}_L q_{HH} \geq \widetilde{q}_{LL}$. Substituting $\widehat{\pi}_L$ gives $\frac{\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})}{q_{HH} - \alpha_L C(q_{HH})} q_{HH} \geq \widetilde{q}_{LL}$, which is equivalent to $\alpha_L \widetilde{q}_{LL} q_{HH} \left( \frac{C(q_{HH})}{q_{HH}} - \frac{C(\widetilde{q}_{LL})}{\widetilde{q}_{LL}} \right) \geq 0$. This inequality holds since $q_{HH} > \widetilde{q}_{LL}$ (because it's assumed that $q_{HH} > \widetilde{q}_{LH}$ and, clearly, $\widetilde{q}_{LH} > \widetilde{q}_{LL}$).

**Uniqueness.** We should guarantee that there is no equilibrium with a pooling component. Consider an equilibrium candidate with a pooling component (i.e. the performance, required with the treat of fine) $\widehat{q} = q_p^*$. Clearly, $q_p^* < \widetilde{q}_{HH}$, since $\widetilde{q}_{HH}$ is the maximal implementable performance under the most favorable beliefs of the Agent and provision of the extrinsic incentive at the same time. If for all $q \leq \widetilde{q}_{HH}$ holds[26] $\pi_L V(q, \alpha_L) < E\left[V(\widetilde{q}_{LL}, \alpha_L)\right]$, then $\alpha_L$-type has a profitable deviation from $\widehat{q} = q_p^*$ to $\widehat{q} = \widetilde{q}_{LL}$ and the equilibrium candidate can't constitute an equilibrium. Taking into account the non-monotonicity of the function $V(q, \cdot)$, the required inequality holds for all $q \leq \widetilde{q}_{HH}$ iff it holds for $q^\times = \min\{\widetilde{q}_{HH}, q^P(\alpha_L)\}$, where $V$ takes the maximal value at $q^P(\alpha_L)$ (see Lemma 1). The inequality $\pi_L V(q^\times, \alpha_L) < E\left[V(\widetilde{q}_{LL}, \alpha_L)\right]$ leads then to[27] $\widehat{\widehat{\pi}}_L = \frac{\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})}{q^\times - \alpha_L C(q^\times)}$ □

## 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 in the decreasing part of the function $U(q; \alpha_L, \beta_H) = \alpha_L \beta_H q - C(q)$ (see Figure 2). Consequently, $\widetilde{q}_{LH} \leq q_{HH}$ is equivalent to $U(\widetilde{q}_{LH}; \alpha_L, \beta_H) \geq U(q_{HH}; \alpha_L, \beta_H)$, leading 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

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

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

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

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 (11), and take the values $\pi_L$ and $\pi_H$, 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

[26] With a slight abuse of notation, $E\left[V(\widetilde{q}_{LL}, \alpha_L)\right]$ denotes here the expected utility when the threat of fine is imposed and performance $\widetilde{q}_{LL}$ is required. The actual performance in this case is $q = \max\{\widetilde{q}_{LL}, q_{LH}\}$.

[27] One can check that for the case of $q_{LH} \geq \widetilde{q}_{LL}$ the threshold is given by $\widehat{\widehat{\pi}}_L = \frac{\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})}{[q^\times - \alpha_L C(q^\times)] - [q_{LH} - \alpha_L C(q_{LH})] + [\widetilde{q}_{LL} - \alpha_L C(\widetilde{q}_{LL})]}$

equilibrium.

□

## Proof of Proposition 2

*Proof.* The existence of the separating equilibrium under the assumptions of the proposition is easy to obtain. In fact, the Agent has correct beliefs, the optimality of the Principal's action follows from the Incentive compatibility conditions, equivalent to the conditions on beliefs, the optimality of the Agent's action follows from Claim 2.

To check the crowding-out condition, consider two cases: $q_c \geq q_{LH}$ and $q_c < q_{LH}$.

If $q_c \geq q_{LH}$ the crowding-out condition is $\pi q_{HH} \geq q_c$. Since $\widehat{\pi}_L \leq \pi \leq \widehat{\pi}_H$, the inequality $\widehat{\pi}_L q_{HH} \geq q_c$ is stronger than the required one. Substituting $\widehat{\pi}_L$ from (19) into $\widehat{\pi}_L q_{HH} \geq q_c$ gives $\frac{q_c - \alpha_L C(q_c)}{q_{HH} - \alpha_L C(q_{HH})} q_{HH} \geq q_c$. After rearranging it leads to $\frac{C(q_c)}{q_c} \leq \frac{C(q_{HH})}{q_{HH}}$, which is equivalent to $q_c \leq q_{HH}$ since the function $C(q)$ is convex. The last inequality is, however, assumed to hold.

If $q_c < q_{LH}$ the crowding-out condition is $\pi q_{HH} \geq \pi q_{LH} + (1-\pi) q_c$, which can be rewritten as $\pi(q_{HH} - q_{LH} + q_c) \geq q_c$. We now prove a stronger inequality $\widehat{\pi}_L(q_{HH} - q_{LH} + q_c) \geq q_c$. Substituting $\widehat{\pi}_L$ from (19) and rearranging leads to

$$q_{HH}\left(\frac{C(q_{HH})}{q_{HH}} - \frac{C(q_c)}{q_c}\right) \geq q_{LH}\left(\frac{C(q_{LH})}{q_{LH}} - \frac{C(q_c)}{q_c}\right)$$

This inequality holds, because $\frac{C(q_{HH})}{q_{HH}} \geq \frac{C(q_{LH})}{q_{LH}}$ since $q_{HH} > q_{LH}$ and $C(q)$ is a convex function.

The inequality $\widehat{\pi}_L > 0$ is evident. The inequality $\widehat{\pi}_H < 1$ for the case of $q_{LH} \leq q_c$ is proven in the same way as for proposition 1. For the case of $q_{LH} > q_c$ the value of $\widehat{\pi}_H$ is given by (18). Notice that $q_{HH} - \alpha_H C(q_{HH}) > q_{LH} - \alpha_H C(q_{LH})$ since the function $q - \alpha_H C(q)$ is increasing for $q \in \left[0, q^A(\alpha_H)\right]$ and $q_{LH} < q_{HH} < q^A(\alpha_H)$ (notice that $C'(q_{ij}) = \alpha_i \beta_j < 1$, whereas $C'(q^A(\alpha_H)) = 1/\alpha_H > 1$). Then, we have denominator in (18) greater than numerator, which establishes $\widehat{\pi}_H < 1$.

To prove uniqueness of the pure-strategy equilibrium, notice that in addition to the fully separating equilibrium[28] it's possible to have pooling equilibria. It's impossible to have pooling on $q = 0$, because then performance $q_{EH} < q_{HH}$ with expected probability $\pi_i$ is realized,[29] giving to $\alpha_L$-type expected utility $\pi_L(q_{EH} = \alpha_L C(q_{EH})) < \pi_L(q_{HH} = \alpha_L C(q_{HH}))$, and $\alpha_L$-type has a profitable deviation to control, according to IC condition (17). It's also impossible to have pooling on control ($q = q_c$), because then $\alpha_H$-type has a profitable deviation to trust ($q = 0$), according to IC condition (16) (we need out-of-equilibrium beliefs to be reasonable in the sense of Cho and Kreps (1987)).

Consider now equilibrium candidates with mixed strategies. In this case, the expected performance in case of control is $q_c$ with probability $\pi_i$ or $q_p^*$ with probability $1 - \pi_i$ (since Agent's beliefs in equilibrium is correct, $q_p^* = \max\{q_c, q^A(E[\alpha|\underline{q} = q_c] \cdot \beta_H\})$, the expected performance in case of no-control is $q_T^*$ with probability $\pi_i$ or 0 with probability $1 - \pi_i$ (the correct beliefs requirement gives $q_T^* = q^A(E[\alpha|\underline{q} = 0] \cdot \beta_H)$. The IC conditions write then as

$$\pi_H(q_T^* - \alpha_H C(q_T^*)) \geq \pi_H(q_p^* - \alpha_H C(q_p^*)) + (1 - \pi_H)(q_c - \alpha_H C(q_c))$$
$$\pi_L(q_T^* - \alpha_L C(q_T^*)) \leq \pi_L(q_p^* - \alpha_L C(q_p^*)) + (1 - \pi_L)(q_c - \alpha_L C(q_c))$$

and can be rewritten as

$$q_T^* - q_P^* + q_c - \alpha_H(C(q_T^*) - C(q_p^*) + C(q_c)) \geq \frac{q_c - \alpha_H C(q_c)}{\pi_H} \tag{23}$$
$$q_T^* - q_P^* + q_c - \alpha_L(C(q_T^*) - C(q_p^*) + C(q_c)) \leq \frac{q_c - \alpha_L C(q_c)}{\pi_L} \tag{24}$$

Consider now three equilibrium candidate profiles:

1) $\alpha_L$-type mixes $q = 0$ and $q = q_c$, $\alpha_H$-type plays pure strategy $q = 0$. In this case $q_p^* = q^\times$, $q_T^* < q_{HH}$. The assumption $\pi \leq \widehat{q}_L$ writes as

$$q_{HH}^* - q^\times + q_c - \alpha_L(C(q_{HH}) - C(q^\times) + C(q_c)) \leq \frac{q_c - \alpha_L C(q_c)}{\pi_L} \tag{25}$$

[28] Clearly, there can be only one fully separating equilibrium. It's impossible to have $\alpha_L$-type choosing $q = 0$ and $\alpha_H$-type choosing $q = q_c$.

[29] I use here notation $q_{EH} = q^A(E[\alpha]\beta_H)$

If $q_{HH}$ is substituted by $q_T^* < q_{HH}$, the inequality in (25) becomes strict because the function $V(q; \alpha_L) = q - \alpha_L C(q)$ is increasing for $q \in [0, q_{HH}]$. On the other hand, the indifference of $\alpha_L$-type between strategies $q = 0$ and $q = q_c$ leads to (24) taken with equality, in contradiction to the just obtained strict inequality.

2) $\alpha_H$-type mixes $q = 0$ and $q = q_c$, $\alpha_L$-type plays pure strategy $q = q_c$. In this case $q_p^* \in [q^\times, q_{EH}]$. The IC condition for $\alpha_H$-type (23) holds with equality. Taking into account that $q_p^* \leq q_{EH}$, and, consequently, $q_p^* - \alpha_H C(q_p^*) \leq q_{EH} - \alpha_H C(q_{EH})$, and substituting $q_{EH}$ instead of $q_p^*$ into (23) leads to

$$q_{HH} - q_{EH} + q_c - \alpha_H(C(q_{HH}) - C(q_{EH}) + C(q_c)) \leq \frac{q_c - \alpha_H C(q_c)}{\pi_H}$$

contradicting to assumption (20).

3) Both $\alpha_L$ and $\alpha_H$-types use mixed strategies. Then $q_p^* > q^\times$ and $q_T^* < q_{HH}$. Both (23) and (24) hold with equality and then should hold

$$(\alpha_H - \alpha_L)\big(C(q_T^*) - C(q_p^*) + C(q_c)\big) = \frac{q_c - \alpha_L C(q_c)}{\pi_L} - \frac{q_c - \alpha_H C(q_c)}{\pi_H}$$

Since $\frac{1}{\pi_L} = \frac{1}{\widehat{\pi}_L} + x$, $\frac{1}{\pi_H} = \frac{1}{\widehat{\pi}_H} - y$ with some $x, y \geq 0$, we get after substituting $\widehat{\pi}_i$ from (18)-(19)

$$\begin{aligned}&(\alpha_H - \alpha_L)\big(C(q_T^*) - C(q_p^*) + C(q_c)\big) = \\ &= (\alpha_H - \alpha_L)\big(C(q_{HH}^*) - C(q^\times) + C(q_c)\big) + x\,(q_c - \alpha_L C(q_c)) + y\,(q_c - \alpha_H C(q_c))\end{aligned}$$

However, the right-hand side of this expression is strictly greater since $C(q_{HH}^*) - C(q^\times) > C(q_T^*) - C(q_p^*)$ and $x, y \geq 0$.

We ruled out the mixed strategy equilibria. Hence, this establishes the uniqueness of the pure strategy separating equilibrium.

□

## Proof of Proposition 3

*Proof.* The proposition is established by checking the equilibrium conditions case by case. Figure 4 illustrates the proof.

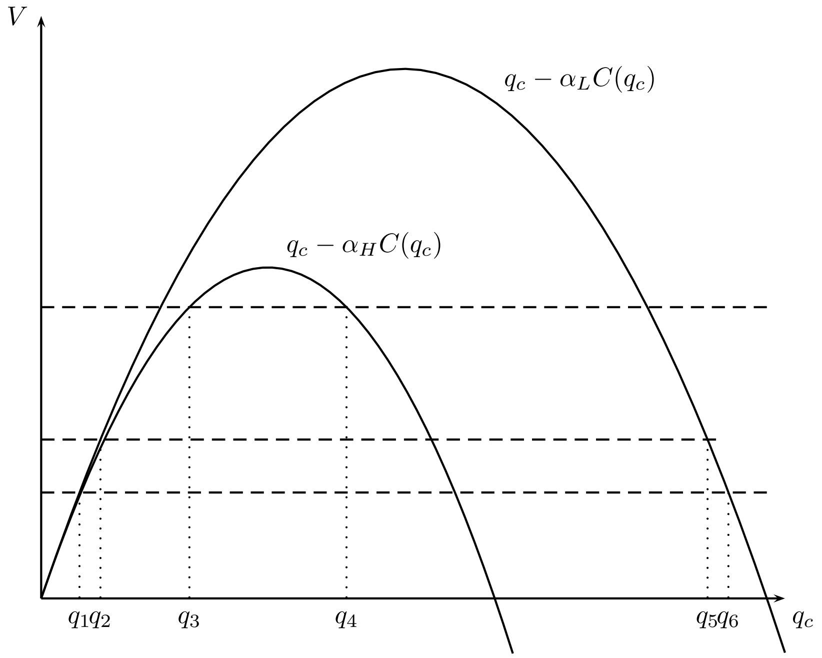

Figure 4: Equilibrium structure in the Control Game

For the no-control pooling equilibrium candidate the optimality (incentive compatibility) conditions for the Principal are written as

$$\pi_H(q_{EH} - \alpha_H C(q_{EH})) \geq \pi_H(q^{\times} - \alpha_H C(q^{\times})) + (1-\pi_H)(q_c - \alpha_H C(q_c)) \tag{26}$$

$$\pi_L(q_{EH} - \alpha_L C(q_{EH})) \geq \pi_L(q^{\times} - \alpha_L C(q^{\times})) + (1-\pi_L)(q_c - \alpha_L C(q_c)) \tag{27}$$

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, determines the threshold $q_1$.

Increasing $q_c$ further leads to the separating equilibrium with crowding-out. The optimality conditions are given by (16), (17). Clearly, (17) taken with equality determines the threshold $q_2$, (16) taken with equality determines $q_3$.

Further increasing $q_c$ leads to control pooling with the optimality conditions

$$\pi_H(q_{HH} - \alpha_H C(q_{HH})) \leq \pi_H(q^{\times\times} - \alpha_H C(q^{\times\times})) + (1-\pi_H)(q_c - \alpha_H C(q_c))$$
$$\pi_L(q_{EH} - \alpha_L C(q_{EH})) \leq \pi_L(q^{\times\times} - \alpha_L C(q^{\times\times})) + (1-\pi_L)(q_c - \alpha_L C(q_c))$$

where $q^{\times\times} = \max\{q_{EH}, q_c\}$. The right-hand side of the optimality condition for $\alpha_H$-type $\pi_H(q^{\times\times} - \alpha_H C(q^{\times\times})) + (1-\pi_H)(q_c - \alpha_H C(q_c))$ is a non-monotone function of $q_c$: it increases for $q \geq q_3$, switching then to decreasing. One can check that this optimality condition is stricter than one for $\alpha_L$-type, and then it determines the two threshold values $q_3$ and $q_4$ (for the threshold $q_4$ one can show that $q^{\times\times} = \max\{q_{EH}, q_c\} = q_c$, which simplifies the right-hand sides of the optimality conditions).

Further increase in $q_c$ leads to the separating equilibrium without crowding-out with optimality conditions, given by (16), (17) with right-hand sides being decreasing functions of $q$. The threshold $q_5$ is determined by (17) taken with equality.

Further increase in $q_c$ leads to an equilibrium involving mixed strategies and then to no-control pooling with optimality conditions (26), (27).

□

# References

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

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

_ **and** _ , "Belief in a Just World and Redistributive Politics," *Quarterly Journal of Economics*, 2006, *121* (2), 699–746.

**_ and _** , “Incentives and Prosocial Behavior,” *American Economic Review*, 2006, *96* (5), 1652–1678.

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

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

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

**Dufwenberg, Martin and George Kirchsteiger**, “A theory of sequential reciprocity,” *Games and Economic Behavior*, 2004, *47* (2), 268–298.

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

**Ellingsen, Tore and Magnus Johannesson**, “Pride and Prejudice: The Human Side of Incentive Theory,” *American Economic Review*, June 2008, *98* (3), 990–1008.

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

**_ and Michael Kosfeld**, “The Hidden Costs of Control,” *American Economic Review*, 2006, *96* (5), 1611–1630.

**_ and Urs Fischbacher**, “A theory of reciprocity,” *Games and Economic Behavior*, 2006, *54* (2), 293–315.

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

**_ and Bettina Rockenbach**, “Detrimental effects of sanctions on human altruism,” *Nature*, 2003, *422*, 137–140.

_ **and Klaus .M. Schmidt**, "The Economics of Fairness, Reciprocity and Altruism–Experimental Evidence and New Theories," *Handbook of the Econonmics of Giving, Altruism and Reciprocity*, pp. 615–692.

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

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

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

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

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

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

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

**Hennig-Schmidt, Heike, 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*, 2005, *55.*

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

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

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

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

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

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

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

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

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

**Sliwka, Dirk**, "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes," *American Economic Review*, 2007, *97* (3), 999–1012.

**Sobel, Joel**, "Interdependent Preferences and Reciprocity," *Journal of Economic Literature*, 2005, *43* (2), 392–436.

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