Miettinen, Topi

Working Paper

Moral hazard and clear conscience

Jena Economic Research Papers, No. 2007,008

Provided in Cooperation with:
Max Planck Institute of Economics

Suggested Citation: Miettinen, Topi (2007) : Moral hazard and clear conscience, Jena Economic Research Papers, No. 2007,008, Friedrich Schiller University Jena and Max Planck Institute of Economics, Jena

This Version is available at:
https://hdl.handle.net/10419/25641

JENA ECONOMIC RESEARCH PAPERS

2007 – 008

Moral Hazard and Clear Conscience

by

Topi Miettinen

www.jenecon.de

ISSN

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich-Schiller-University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact m.pasche@wiwi.uni-jena.de.

Impressum:

Friedrich-Schiller-University Jena
Carl-Zeiß-Str. 3
D-07743 Jena
www.uni-jena.de

Max-Planck-Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Jena Economic Research Papers 2007-008

Moral Hazard and Clear Conscience

Topi Miettinen
Max Planck Institute of Economics*
RUESG, University of Helsinki

April 17, 2007

Abstract

We consider guilt averse agents and principals and study the effects of guilt on optimal behavior of the principal and the agent in a moral hazard model.

The principal's contract proposal contains a target effort in addition to the monetary incentive scheme. By accepting the agreement, the parties agree on both the wage scheme and the target.

The agent suffers from guilt when failing to provide the target effort, the principal when paying less than the contract requires or when setting an unreasonably high target effort.

In equilibrium, a guilt-prone agent chooses a higher effort than an agent who only cares about monetary incentives. The target effort level is always set above the equilibrium effort. Both the agent and the principal gain from the agent's guilt aversion.

A principal who lacks power to commit to the proposed incentive scheme benefits from having a positive proneness to guilt. However, a guilt-prone principal who suffers when setting an unreasonable target is worse off than one with merely monetary motivations.

KEYWORDS: Moral Hazard, Norms, Agency, Social Preferences
JEL: C72, D82, Z13

*Kahlaische Straße 10, 07745 Jena, Germany; miettinen@econ.mpg.de

1 Introduction

"Then I thought a minute and says to myself: Hold'on. S'pose you'd done right and give Jim up. Would you felt better than you do now? No, says I. I'd feel just the same I feel now. Well then, I says. What's the use of learning to do the right when doing right is troublesome and there is no trouble doing wrong and the wages is just the same. I was stuck. I could not answer that. So I reckoned. I would no more bother about it but after this do always whichever comes handiest at time." (Mark Twain: Huckleberry Finn).

The quotation above was put forward by Holmström and Milgrom (1987) to highlight Huckleberry's rational reasoning that leads him to choose an action that is the best in terms of wages and ease of use. It leads us to the roots of the problem of providing incentives to a risk-averse agent: paying more when the output is high provides incentives to work hard but if output depends on other factors than agent's effort, a risk-averse agent must be compensated for accepting the risk. Yet, the quotation also highlights Huckleberry's trade off between choosing the right against choosing 'whichever comes handiest at time'. Huckleberry reasons that when both doing right and doing wrong make him feel equally good about himself, the best choice is the one that takes least effort.

Huckleberry goes further and asks himself: 'What's the use of doing the right...?' In other words, can Huckleberry gain from feeling better about doing the 'right'? Huckleberry reasons that the answer must be negative: preference for choosing the right only prevents him from choosing the handiest at time and, hence, such preferences cannot pay off.

We illustrate in the single dimensional moral hazard model (Holmström, 1979) that Huckleberry's answer may be incorrect: when the preference for doing right is observed by the principal, it provides commitment power. If Huckleberry is known to prefer to do as agreed, an agreement on how much effort Huckleberry should provide is no longer cheap talk and can be used as a riskless alternative to a high-powered incentive scheme to induce effort. We show that an agent who feels bad about doing wrong is paid higher wages in equilibrium, even if her monetary incentives to provide effort are lower.

Formally, we introduce two additional features into Holmström (1979). First, the contract offer made by the principal includes an explicit target effort level in addition to the monetary incentive scheme. In the standard model, this target effort is restricted to coincide with the agent's optimal effort choice. The novelty in our model is that the target does not have to coincide with the agent's actual effort. The second new ingredient of our model is that the agent may have a preference for clear conscience: she suffers a cost if she fails to meet the target agreed in the contract.

Surprisingly, using an explicit effort level as an incentive is not entirely costless for the principal, however. The optimal agreement asks for an unoptimally high effort from Huckleberry (from his perspective) and the bad feelings about not meeting the target must be compensated for by the principal so that Huckeberry is willing to accept the contract. However, since the adopted incentive scheme is less risky than one which does not take advantage of Huckeberry's

moral preferences, both Huckleberry and his employer get higher earnings than if Huckleberry felt equally good about doing right and doing wrong.

We also consider the case where, in addition to the agent, also the principal is motivated by preference for doing right. Once the output has been created, it is in the interest of the principal not to pay the agent but rather to keep the entire output to herself. We first consider such a scenario and illustrate that a principal with observable preference for doing right is better off than a principal without since the latter cannot commit to pay and, therefore, the agent provides no effort or rejects the offer. We also consider another scenario where the principal feels bad if she sets an unrealistic target for the agent. In this case a principal without any concern for doing right is better off than one with such a preference.

A bulk of literature considers the effects of agent's equity concerns on optimal contracts. The models closest to our setup are those of Englmaier and Wambach (2005), Itoh (2004), Dur and Glazear (2004) who consider an agent who compares her payoff to that of the principal[1]. Unlike the current paper, all these models essentially do not extend the general model of Holmström (1979) but rather simplify by assuming a risk-neutral and/or contractible effort. They all find that the principal's equilibrium payoff decreases and that of the agent increases if the agent is more concerned about equity: while when failing to produce output, the agent can be paid according to her outside option compensation, an envious agent must be paid more in the case of success to make sure that the principal does not get too large a share of gross profits. This paper illustrates how plausible other-regarding preferences may have quite the opposite effect on optimal contracts: a lower powered incentive scheme and gains both to the agent and to the principal from the agent being more other-regarding.

The paper most related to ours is Akerlof and Kranton (2005). There, the principal may take measures to make the agent identify more strongly with the firm and its goals. The identity in their model functions like the target effort in our model since the identity is essentially a preference for doing as the identity calls for. As in the present paper, the induced target provides an alternative to the high powered incentive scheme to induce effort. The model in Akerlof and Kranton (2005), however, completely abstracts from the endogenous cost of inducing a higher target effort: the bad feelings of not reaching the target must be compensated for. Rather, they assume an exogenous cost of building up firm identity. Apart from the exogenous costs/benefits of firm identity, the present model can be considered as a generalization of Akerlof and Kranton (2005) which illustrates that the principal faces a trade-off when using informal normative targets even when using the informal target is not directly costly.

The paper is organized as follows. Section 3 presents the general moral hazard model with agent having a preference for clear conscience. An example in the linear normal setup provides some further intuition. Section 4 considers a principal with proneness to guilt. Section 5 concludes.

[1] Papers that consider the effects of social preferences on optimal contracts in team production include Biel-Rey (2002), Huck et al. (2006), Rob and Zemsky (2002).

2 Approaches to other-regarding preferences

On the one hand, economics has been successful in incorporating and capturing essential features and properties of social behavior such as reputation effects and punishment strategies in repeated games. On the other hand, it has until recently, ignored the fact that people value behavior that conforms with social ideals intrinsically, independently of its material payoff.

Among laymen, the trade off between moral and material payoffs is probably the most discussed class of economic problems on the planet. Not choosing the moral ideal causes guilt and reduces payoff. This idea is not new in the realm of economics. The literature on social preferences[2] provides with numerous counterexamples: Rabin (1993), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Dufwenberg and Kirchsteiger (2004). Not incorporating social ideals explicitly into the players' preferences implies that a descriptive theory may fail to capture some essential features which may sometimes hinder plausible results, predictions and policy implications.

However, introducing social preferences explicitly into agent's preferences has aroused some objections. They seem to offer a much too flexible a tool to explain any behavior observed just by choosing some social preferences that fit the observed behavior. But similarly in a descriptive model, considering only material payoffs is itself a restrictive, even if simplifying, hypothesis which has to be justified. Then the question is, on what basis the properties of social preferences should be determined. There are several ways to give arguments in favour of a family of social utility functions. Introspection is a first criterion. Experiments and empirical research is another. Evolutionary preference models is a third.

In the spirit of the evolutionary argument, we can ask whether a player can gain in terms of material self interest from having a preference for not choosing according to material self-interest? The indirect evolutionary literature (Güth and Yaari, 1992; Samuelson, 2001) has shown that for some games and for some suitably tailored preferences this is indeed the case. Utility function that is not maximized at a point where the material self interest reaches its maximum gives a player a means of committing to actions that otherwise would not be credibly chosen. Consequently, the opponent may alter behavior and this may increase material payoff.

Nevertheless, it is not evident that the fairness or the reciprocity approaches provide simple and unified explanations to capture the social concerns in a wide variety of economic interactions. In ethics and in welfare economics, there is a long-lasting and still ongoing dispute about which principles of justice should be applied. As among philosophers and welfare economists, also among laymen there are proponents of each moral principle - there is potentially a large number of various principles that can be internalized and formalized as a social utility function. The concern for equity is just one particular principle of distributional justice. Similarly, reciprocity could be thought as a concern to contribute to the

[2] See Sobel (2005) for a review.

moral principle of being nice to those who are nice to you and punish those who do not. With so many ethical principles fighting for popularity across agents and across interaction contexts, there is little hope that a single principle, such as one of those formulated in Rabin (1993), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Dufwenberg and Kirchsteiger (2004) will provide the ultimate salient explanation[3].

In game theoretic terms, that a norm is common knowledge seems like a good way to formalize saliency. Indeed, social psychologists argue that for a normative ideal to have bite, it should be shared among all agents involved Millar and Tesser (1988). There are two likely sources of violation of this common knowledge condition both in the reciprocity approach and in the inequity aversion approach. First, each player's preferred general moral principle tends to be private information and even the support of potential principles is hard to imagine, let alone the support or the distribution of the ideal actions implied by these principles. Second, even if the principle and thus the social preference is common knowledge, a descriptive model should take into account that there may not be common knowledge about the ability of all parties involved to infer the implied normative action from applying the general principle to the potentially complex interaction context.

However, the first problem does not exist and the second is alleviated if the socially ideal actions themselves are common knowledge. An action-norm is a joint plan of action capturing the socially ideal behavior. If there is common knowledge about an action-norm, then all parties know that all parties know (ad infinitum) how each party should normatively behave. Thus, the task of verifying whether the action-norm is an equilibrium becomes much simpler when the norm itself does not have to be inferred first from the context.

How are such action-norms established? This paper considers them as an output of pre-play communication/negotiation[4]. When there is an action-norm in place, there is a shared expectation of the normatively ideal behavior. Social psychologists argue that in such a situation people tend to feel guilty[5] about

[3] See Engelmann and Strobel (2004), Charness and Rabin (2002), Charness and Dufwenberg (2006) for evidence on concerns for ethical principles not accounted in fairness and reciprocity models.

[4] There are two important forums of pre-play negotiation. First, when all parties jointly come together to negotiate a joint plan before the game is played. Second, when all parties involved are commonly known to be members of a given community with a shared (sub)culture. In this latter case, the parties involved may commonly know that the community has reached a consensus of an action-norm in this game in its grand scale pre-play negotiation (see (**?**, Akerlof and Kramton)). This latter type of pre-play negotiation is the moral discourse or the moral gossiping within the community. In the moral discourse, people discuss conflicts and violations of norms and present arguments for and against adopted actions (using general principles of justice, for instance). Every time this moral discourse occurs, the parties may or may not reach a consensus. When the discourse over a particular game is repeated over time and when sequence of outcomes of such discourses (say in the presence of any subgroups of a community) comes to a rest point - that is, each negotiation concludes with the same joint action profile - common knowledge of an action-norm is established. For a more general model, see Miettinen (2006).

[5] Dufwenberg and Gneezy (2000), Dufwenberg (2002), Charness and Dufwenberg (2006), Miettinen (2006) consider simple models of guilt which have the action-norm interpretation.

violating the norm and, more importantly, that the guilt is increasing in the harm that the violating causes to others. Therefore, if players can observe each others' preferences, they can verify whether the ideal action profile is an equilibrium, that is whether other players would feel sufficiently guilty about deviating from the ideal to prevent them to deviate.

3 Agent with Preference for Clear Conscience

3.1 The general case

Let us consider a single-dimensional moral hazard model, Holmström (1979). Risk-neutral principal owns a stochastic production technology. The output level is denoted $q \in (-\infty, \infty)$. Principal hires an agent to control the technology and proposes an incentive scheme $s(q)$ to the agent. Thus the expected payoff of the principal is

$$\int_{-\infty}^{\infty} (q - s(q)) f(q; a) dq,$$

where $f(q; a)$ is the density function of the output q. Agent chooses an action $a \in [0, \infty)$ and output is drawn randomly from a distribution that is parametrized with the action. The cumulative distribution function is $F(q; a)$. We suppose that $F_a(q; a) \leq 0$ and that for every $a' > a''$, $F_a(q; a') < F_a(q; a'')$ so that $F_a(q; a')$ first order stochastically dominates $F_a(q; a'')$.

Von Neumann-Morgernstern utility function $u(s(q))$ is agent's material gross payoff. Agent is strictly risk averse

$$u'' < 0 < u'.$$

Agent's utility is separable in money and effort. The agent suffers a physical cost of effort captured by function $c(a) : R_+ \longrightarrow R_+$ which is increasing and convex on the set of actions and $c(0) = 0$

$$c', c'' > 0.$$

The agent has preference for clear conscience. She suffers from guilt if she inflicts harm on the principal by providing less effort than agreed. We assume a very specific form of the clear conscience utility function where $g = \frac{1}{2}(\min\{a - a^*, 0\})^2$ so that implicitly the agents suffers only if she harms the principal by violating the norm a^* and the guilt is increasing in the harm inflicted on the principal[6]. Hence, the cost function of an agent with clear conscience preferences then can be written as

$$C(a, a^*; \delta) = c(a) + \frac{\delta}{2}(\min\{a - a^*, 0\})^2.$$

[6] For the sake of simplicity, we do not make guilt a function of the expected harm of the principal explicitly.

where $\delta \in [0, \infty)$ is agent's proneness to guilt. We assume that the agent's preferences and the physical cost is observable to the principal. We discuss the alternative assumptions in the conclusion.

The game is structured as follows: prior to the agent's choice, the parties negotiate. The principal makes a take-it or leave-it proposal with two items to the agent: monetary incentive scheme $s(q)$ and a target action a^*. Agent can either accept or reject the contract. If an agent with a positive proneness to guilt accepts the proposal and deviates from target effort, she will suffer from guilty conscience. The agent has an outside option $\underline{u}$ which captures the opportunity cost of the agent.

The optimization problem of the principal is written as[7]

$$\max_{a,\, a^*, s(q)} \int_{-\infty}^{\infty} (q - s(q)) f(q; a) dq$$

$$s.t.$$

$$\int u(s(q)) f_a(q; a) dq - \delta(a - a^*) - c'(a) = 0 \tag{1}$$

$$\int u(s(q)) f(q; a) dq - \frac{\delta}{2}(a - a^*)^2 - c(a) \geq \underline{u} \tag{2}$$

Proposition 1 *The optimal incentive scheme is implicitly characterized by* $\frac{1}{u'(s(q))} = \lambda + \mu \frac{f_a}{f}$. *The coefficient* μ *is positive. Effort level chosen by the agent is below the target effort level. Optimal target effort level is given by* $a^* = a + \frac{\mu}{\lambda}$. *As* δ *tends to infinity,* $a^* - a$ *tends to zero.*

Proof. The first-order conditions of the Lagrangian w.r.t. a^*, s and a are given by

$$\frac{\partial L}{\partial a^*} = \mu\delta + \lambda\delta(a - a^*) = 0 \tag{3}$$

$$\frac{\partial L}{\partial s} = -f(q; a) + \mu[u'(s(q)) f_a(q; a)] + \lambda[u'(s(q)) f(q; a)] = 0 \tag{4}$$

$$\frac{\partial L}{\partial a} = \int (q - s(q)) f_a(q; a) dq + \mu\{\int u(s(q)) f_{aa}(q; a) dq - \delta - c''(a)\} = 0 \tag{5}$$

Then from (3), it follows that

$$a^* = \frac{\mu + \lambda a}{\lambda} = a + \frac{\mu}{\lambda} \tag{6}$$

[7] The first order approach assumes that the solution to the agent's maximization problem is given by the effort which render the first derivative of the target function zero. Jewitt (1988) provides sufficient conditions.

which is greater than a when μ is positive. And from (4), it follows that

$$\frac{1}{u'(s(q))} = \lambda + \mu\frac{f_a}{f} \tag{7}$$

The latter is a result analogous to that in Holmström (1979) and it states that monetary reward is increasing in output provided that λ and μ are positive. To show that coefficient μ is positive, follow the steps in lemma 1 in Jewitt (1988). From (1)

$$\int u f_a(q;a)dq = c'(a) + \delta(a - a^*) \tag{8}$$

(7) gives

$$f_a = f(\frac{1}{u'} - \lambda)\frac{1}{\mu}$$

Plugging this into (8) gives

$$\int u(\frac{1}{u'} - \lambda)\frac{1}{\mu}f(q;a)dq = c'(a) + \delta(a - a^*)$$

Taking the expectation on both sides of (7) gives

$$E[\frac{1}{u'(s(q))}] = \lambda$$

Then

$$\int u(\frac{1}{u'} - \lambda)f(q;a)dq$$

has sign of covariance of $\frac{1}{u'}$ and u. This sign is positive. Therefore, μ takes the sign of $c'(a) + \delta(a - a^*)$. That is

$$sgn(\mu) = sgn(c'(a) + \delta(a - a^*))$$

In addition, from (6), we get

$$sgn(\mu) = sgn(c'(a) - \delta\frac{\mu}{\lambda}) \tag{9a}$$

Hence μ cannot be non-positive because with non-positive μ, the right hand side of (9a) is strictly positive and the equality does not hold. Hence, μ must be strictly positive. This implies that

$$a^* > a.$$

Moreover, μ is given by the solution to (5). Therefore, it is straightforward to see that as δ tends to infinity, μ tends to zero. Thus, from (6), we get that $a^* - a$ tends to zero as δ tends to infinity. ■

The intuition behind this result is simple. High-powered monetary incentives and target effort are substitutes in inducing effort. High powered monetary incentives imply a cost, because the agent is risk-averse and does not want the

income to be tied on the stochastic output. To reduce marginal cost of inducing effort, the principal has an incentive to make monetary incentives lower-powered and to use the target effort instead. The principal sets the target effort above the commonly known equilibrium effort. This creates tensions between moral and material optima in the agent's decision problem. The agent trades off the moral and the material payoffs and, in equilibrium, she suffers from guilty conscience. This creates an indirect cost to the principal since the agent must be compensated for the bad feelings. In equilibrium, the marginal disutility of guilty conscience equals the marginal disutility of bearing risk. The principal gains because each level of effort can be implemented with a lower cost. On the other hand, when the agent has an infinite proneness to guilt, she will not deviate from the target effort level and thus does not suffer from guilt. The target effort is set to coincide with the first-best effort. The agent gets a fixed remuneration which barely guarantees that the agent accepts the offer.

The following proposition shows that the agent with a positive proneness to guilt will reach higher earnings than a zero proneness counterpart even if the monetary incentives of the latter may be higher powered. This may be surprising at first sight. One might conjecture that, since weaker monetary incentives induce the same or higher effort, the agent would seem to lose from a positive proneness to guilt. However, the principal pays the agent the lowest remuneration that she still accepts. All types have equal wages in an outside option[8]. The result above shows that in equilibrium the agent prone to guilt suffers since she never reaches the target effort. The agent must be compensated also for having to feel guilt to make her accept the job in the first place. Hence, the material payoff of a guilt-prone agent will be above the outside option payoff. What is more, it does not pay off for the agent to have an infinite proneness to follow the contract, because then the marginal cost of using clear conscience incentives is infinite and this device will not be used. Agents with an intermediate proneness to guilt get better material payoffs.

Proposition 2 *Expected material payoff for the agent with proneness to guilt $\delta \in (0, \infty)$ is higher than the material payoff of the agent with zero proneness to guilt and agent with infinite proneness to guilt.*

Proof. It is easy to see that in equilibrium agent's material payoff

$$\begin{aligned}\int u(s_\delta(q))f(q;a_\delta)dq - c(a_\delta) &> \int (s_\delta(q))f(q;a_\delta)dq \\ &\quad -\frac{\delta}{2}(a_\delta - a^*)^2 - c(a_\delta) \\ &= \underline{u} \\ &= \int u(s(q))f(q;a)dq - c(a),\end{aligned}$$

[8] A job where effort is perfectly monitored for instance.

where a_δ and a are the effort levels chosen by agent with proneness δ to clear conscience and zero proneness to guilt respectively. Hence, an agent with positive δ earns strictly more in expectation than agent with $\delta = 0$. The latter result follows from the finding in the previous proposition that $a_\delta - a^*$ approaches zero as δ tends to infinity. ■

Corollary 3 *The expected payoff of the principal and the social surplus are higher when principal faces an agent with* $\delta > 0$.

Proof. For each positive proneness to guilt, the principal could propose the agent $a^* = a_{\delta=0}$ the incentive scheme $s_{\delta=0}(q)$ and get exactly the same payoff for each δ. However it is shown above that such a policy is not optimal when there is a positive proneness to guilt. It is also shown that the agent with $\delta > 0$ gets a higher payoff than an agent with $\delta = 0$. The sum of payoffs is then greater as well. ■

We saw above that it is profitable for the principal to use target effort to induce effort because it reduces cost of implementing inframarginal units of effort. Optimality with a risk neutral principal requires that the expected marginal productivity of effort equals its marginal cost. The effort level will be higher than in a model without proneness to guilt, which further increases overall welfare.

Provide monetary incentive schemes that condition pay on output or profit alleviates the agency problem between the owner of a production technology and the agent she hires. However, the empirically observed monetary incentives are often lower powered than theory predicts. People are paid a somewhat fixed remuneration and some targets are set despite the fact that realized action is not observable or enforceable. Existence of clear conscience preferences may be one explanation.

However, counting too heavily on moral incentives may be naive. Section 4 shows that for the principal it is better not to have a positive proneness to guilt. Managers act as agents towards the firm owners but as principals towards other workers of the firm and , therefore, being guilt averse may not be universally beneficial.

3.2 Linear-Normal example

In this section we consider a model where the incentive scheme is restricted to a linear one and where the agent controls the mean of a normally distributed output the variance of which does not depend on the action. In this simple case, the optimal solution can be solved for explicitly and thus we use it to better illustrate the intuition of the model[9]. The assumptions of the model are as follows:

(a) the output is normally distributed with $q \tilde{} N(a, \sigma^2)$

[9] Holmström and Milgrom (1987) motivate this approach by showing that it is a reduced form of a problem of incentivizing the agent who must control a technology over a longer time interval.

(b) the incentive scheme is linear $s(q) = vq + t$

(c) the cost of effort is written as $c(a) = \frac{c}{2}a^2$

(d) the agent's utility has a constant absolute risk aversion $u(y) = -\exp(-ry)$.

where r is the coefficient of absolute risk aversion and $y = vq + t - \frac{c}{2}a^2 - \frac{\delta}{2}(a - a^*)^2$. These assumptions allow us to derive the equilibrium strategies in a reduced form and to gain some further intuition.

We can write the principal's maximization problem as follows:

$$\max_{v,t,a^*} \int \{a + \varepsilon - v(a + \varepsilon) - t\}h(\varepsilon)d\varepsilon$$

$$s.t.$$

$$\int \{-\exp[-r(v(a + \varepsilon) + t - \frac{c}{2}a^2 - \frac{\delta}{2}(a - a^*)^2)]\}h(\varepsilon)d\varepsilon \geq \underline{u}$$

$$\arg\max_{a} \int \{-\exp[-r(v(a + \varepsilon) + t - \frac{c}{2}a^2 - \frac{\delta}{2}(a - a^*)^2)]\}h(\varepsilon)d\varepsilon$$

where $\varepsilon \tilde{} N(0, \sigma^2)$ and $h(.)$ is the density of this normal distribution. The problem can be alternatively written[10] as

$$\max_{v,t,a^*} \{(1 - v)a - t\} \tag{10}$$

$$s.t.$$

$$a = \frac{v + \delta a^*}{c + \delta} \tag{11}$$

$$t = \underline{u} - vq + \frac{rv^2\sigma^2}{2} + \frac{\delta}{2}(a - a^*)^2 + \frac{c}{2}a^2 \tag{12}$$

where (11) is the solution to

$$\max_{a} \{va + t - \frac{rv\sigma^2}{2} - \frac{c}{2}a^2 - \frac{\delta}{2}(a - a^*)^2\}.$$

This latter is equivalent to the agent's maximization problem. From (11), it is now easy to see that the monetary incentives and the target effort are substitutes in inducing effort. Moreover, an agent more prone to guilt is less responsive to monetary incentives than a standard agent.

[10] See page 138 in Bolton and Dewatripont (2005) for details in the case where the agent has no explicit preferecne for doing right.

It is easy to check that the second order condition of the problem is satisfied.

Plugging (11) and (12) into (10) and maximizing, we get the optimal bonus rate,

$$v_\delta = \frac{1}{(1+(c+\delta)r\sigma^2)}, \tag{13}$$

and the optimal target effort,

$$a^*_\delta = \frac{1}{c}. \tag{14}$$

Remark 4 *Independently of the agent's type, the principal sets the target effort equal to the first best effort of the agent*[11]*. The agent prone to guilt is offered a lower bonus rate than the agent with zero proneness to guilt.*

Even a guilt-prone agent never provides the first best effort.

Remark 5 *The target effort is always above the equilibrium effort:*

$$a^*_\delta - a_\delta = \frac{r\sigma^2}{(1+(c+\delta)r\sigma^2)} > 0. \tag{15}$$

Plugging (13) and (14) into (12) and recalling that $q = a + \varepsilon$ we get

$$t_\delta = \underline{u} + \frac{(rc\sigma^2 - 1) + \delta(c+\delta)(r\sigma^2)^2}{2c(1+(c+\delta)r\sigma^2)^2} \tag{16}$$

Remark 6 *Fixed remuneration for the guilt prone agent is higher and the risk neutral principal bears a larger share of the risk which improves efficiency.*

An agent with a positive proneness to guilt, $\delta > 0$, chooses

$$a_\delta = \frac{c + \delta(1+(c+\delta)r\sigma^2)}{c(c+\delta)(1+(c+\delta)r\sigma^2)} \tag{17}$$

and an agent with zero positive proneness to guilt chooses

$$a = \frac{1}{c(1+rc\sigma^2)}.$$

Remark 7 *The guilt-prone agent chooses a higher effort level than the agent without proneness to guilt.*

The agent's certainty equivalent must exceed the payoff of her best outside option. In optimum, the certainty equivalent is equal to the outside option payoff:

$$v_\delta a_\delta + t_\delta - \frac{r v_\delta^2 \sigma^2}{2} - \frac{c}{2}a_\delta^2 - \frac{\delta}{2}(a_\delta - a^*_\delta)^2 = \underline{u}.$$

[11] The first-best is given by $arg\max_a E(q|a) - c(a)$, or equivalently $a = \frac{1}{c}$.

Notice that $v_\delta a_\delta + t_\delta - \frac{rv_\delta^2\sigma^2}{2} - \frac{c}{2}a_\delta^2$ is the agent's material payoff and thus the difference between the payoffs of a guilt-prone agent and a standard agent is merely $\Pi(\delta) \doteq \frac{\delta}{2}(a_\delta - a_\delta^*)^2$. Plugging (14) and (17) into $\frac{\delta}{2}(a_\delta - a_\delta^*)^2$ we obtain

$$\Pi(\delta) = \frac{\delta(r\sigma^2)^2}{2(1+(c+\delta)r\sigma^2)^2}.$$

This term is the difference in material payoffs of an agent with proneness δ to clear conscience and an agent with zero proneness to guilt. By definition,

$$\Pi(0) = 0.$$

On the other hand, applying Hospital's rule gives

$$\lim_{\delta\to\infty} \Pi(\delta) = 0.$$

Remark 8 *The agent who is infinitely prone to guilt is equally well of as an agent with zero proneness to guilt.*

It is easy to see from (15) that the agent who is infinitely prone to guilt provides the first best effort and gets a flat compensation. All rents accrue to the principal.

To see that there exists an intermediate proneness to guilt that does better, consider $\Pi'(\delta)$ which is continuous and positive when δ is zero and $sgn(\Pi'(\delta)) = sgn(1 + cr\sigma^2 - \delta r\sigma^2)$.

Remark 9 *There is a unique value,*

$$\delta = \frac{1+cr\sigma^2}{r\sigma^2} > 0,$$

which maximizes the material payoff of the agent.

4 Principal with Preference for Clear Conscience

In this section, we consider two alternative scenarios where both the principal and the agent may suffer from guilt. In the first scenario, the principal does not have any exogenous commitment device that guarantees that she will ex post pay according to the contract that she offers ex ante. Instead, the agent may be held up and paid less than agreed when the output is realized and the payment is due. Naturally, guilt about not paying as agreed provides the principal with an intrinsic partial commitment device if this preference is observed by the agent ex ante when the contract is offered. In the second scenario, the principal has access to an exogenous commitment device, but she may suffer from setting the target effort at an unreasonable level for the agent (what is meant by unreasonable is to be defined below). The principal's and the agent's proneness to guilt are denoted by δ_P and δ_A, respectively.

4.1 Lack of Commitment

Let us now consider the first scenario. The principal suffers from guilt if she pays less than the amount indicated in the incentive scheme, $s(q)$. Let us denote the actual payment as a function of output by $t(q)$. When the uncertainty is resolved and the output is realized, the principal needs to decide how much to pay the agent given the output and the incentive scheme that was agreed upon, $q - t(q) - \frac{\delta_P}{2}(s(q) - t(q))^2$ where $q - t(q)$ is the material payoff and $\frac{\delta_P}{2}(s(q) - t(q))^2$ is the principal's guilt cost. An interior solution to the problem is

$$t(q) = s(q) - \frac{1}{\delta_P}.$$

The agent perfectly anticipates the lack of commitment of the principal. Thus, the principal can implement the original scheme by setting $s(q) = t(q) + \frac{1}{\delta_P}$ where $t(q)$ now equals the original scheme. The principal gets exactly the same expected material payoff as before, but now in addition, she suffers the cost of breaching $-\frac{1}{2\delta_P}$. Notice yet, that this procedure is out of reach of the principal with zero proneness to guilt. The agent correctly anticipates that the principal will not pay anything in any case. So the agent will not put in any effort and chooses her outside option $\underline{u}$.

Proposition 10 *Without an exogenous commitment device, the principal with $\delta_P = 0$ is worse off than one with any $\delta_P > 0$. The expected material payoff is the the same for all $\delta_P \in (0, \infty)$.*

4.2 Aversion for an unreasonable target

Suppose now instead that the principal has full commitment power. The agent does not have to worry about threat of a hold-up. This may be due to institutional reasons such as legal enforcement of the contract. Let us however suppose that the principal prefers not setting an unreasonably high target effort to the agent. In other words, the principal suffers from guilt if the actual effort and the target effort are not equal in equilibrium. The principal knows that the agent is going to choose

$$a = \frac{v + \delta_A a^*}{c + \delta_A}$$

Thus, the principal's morally ideal target effort is the only fixed point of the above mapping, i.e.

$$a^* = \frac{v}{c} \tag{18}$$

A principal with a positive proneness to guilt, has a higher cost of using a target effort level as a means of inducing effort. Thus, the equilibrium target and the actual equilibrium effort are closer to each other and monetary incentives are higher-powered than those of a principal with no proneness to guilt. We

confirm these intuitions below in the linear-normal model. We also show that the material payoff of the principal with a positive proneness to guilt is lower payoff than that of with a zero proneness to guilt.

Given variable y, let y_0, y_δ and y_Δ respectively denote the equilibrium values of the variable in cases with (i) no proneness to guilt, (ii) agent with a positive proneness to guilt only and (iii) agent and principal with a positive proneness to guilt.

Proposition 11 *Let assumptions (a)-(d) hold. When principal has a positive proneness to guilt* $v_\delta < v_\Delta < v_0,\ t_\delta > t_\Delta > t,\ a_\delta > a_\Delta > a,\ a^*_\delta > a^*_\Delta > a^*,\ (a_\delta - a^*_\Delta) > (a_\Delta - a^*_\Delta) > (a_0 - a^*_0) = 0.$

Principal with a positive proneness to guilt is worse off than a principal with no proneness to guilt.

5 Discussion

Clear conscience preferences are introduced into a hidden action moral hazard setup. The effects of clear conscience preferences on equilibrium behavior of the principal and the agent are studied. When facing an agent prone to guilt, the principal sets target action above the equilibrium effort choice of the agent and uses guilt to induce effort. Hence in equilibrium, the agent suffers from guilt. Because the agent must be compensated sufficiently to accept the job in the first place, an agent prone to guilt receives a higher fixed remuneration and, therefore, higher earnings than an agent with zero proneness to guilt. However, infinite proneness to guilt does not pay off, because the agent will never deviate from the target effort and thus there no need to compensate for the bad feelings. Infinite proneness to guilt maximizes welfare however, since the agent chooses the first best action.

A principal who is prone to guilt receives higher earnings than one not prone to guilt when there is no exogenous device that commits the principal to her contract offer. An agent who fears being held up and paid less than agreed will not accept the contract.

Yet, if the principal can credibly commit without intrinsic motivation, a principal cannot gain from being prone to guilt. As an example, we discuss the case where a principal dislikes setting targets at an unreasonably high level.

6 Appendix

Proof of proposition 11. Principals problem can be written as

$$\max_{a,a^*,v}\{a - \frac{rv\sigma^2}{2} - \frac{c}{2}a^2 - \frac{(\delta_A + \delta_P)}{2}(a - a^*)^2 - \underline{u}\}$$

$$s.t. \tag{19}$$

$$a = \frac{v + \delta a^*}{c + \delta}$$

Redefine $x = a - a^*$. Then $a = a^* + x$. Plug in new variables and rewrite the problem:

$$\max_{a,a^*,x} \{a^* + x - \frac{rv\sigma^2}{2} - \frac{c}{2}(a^* + x)^2 - \frac{(\delta_A + \delta_P)}{2}x^2 - \underline{u} - \mu(a^* + x - \frac{v + \delta a^*}{c + \delta})\}$$

First-order conditions:

$$a^* = \frac{1 - cx - \mu(\frac{c}{c+\delta_A})}{c} = 0 \tag{20}$$

$$v = \mu \frac{1}{(c + \delta_A)r\sigma^2} \tag{21}$$

$$1 - c(x + a^*) - \mu - (\delta_P + \delta_A)x = 0 \tag{22}$$

$$a^* + x = \frac{v + \delta a^*}{c + \delta} \tag{23}$$

Solving with respect to parameters gives

$$x = -\frac{\delta_A}{(c + \delta_A)[(\delta_P + \delta_A)(1 + cr\sigma^2) + (\delta_A)^2 r\sigma^2]} \tag{24}$$

$$\mu = \frac{(\delta_A^2 + \delta_A + \delta_P)(1 + cr\sigma^2)}{c} \tag{25}$$

Hence, we learn that

$$\frac{\partial x}{\partial \delta_P} > 0$$

but by proof above $x < 0$. Hence the absolute value of x is decreasing. Then

$$\frac{\partial v}{\partial \delta_P} > 0$$

$$\frac{\partial a^*}{\partial \delta_P} < 0$$

$$\frac{\partial a}{\partial \delta_P} < 0$$

Summing up: target effort level is set below, incentives are higher powered, equilibrium effort is lower, fixed remuneration is lower than without principal's proneness to guilt. Define $\Pi(\delta_A, \delta_P) \equiv \frac{\delta_A}{2}(a(\delta_A, \delta_P) - a^*(\delta_A, \delta_P))^2$. Clearly G is continuous and

$$\Pi(\delta_A, 0) = \frac{\delta_A (r\sigma^2)^2}{2(1 + (c + \delta_A)r\sigma^2)^2}. \tag{26}$$

Applying l'Hospital's rule

$$\lim_{\delta_A \to \infty} \Pi(\delta_A, \delta_P) = 0 \tag{27}$$

In addition, with some simple calculations one can show that there exists a unique point where

$$\Pi_1(0, \delta_P) = 0$$

To sum up $\Pi(\delta_A, \delta_P)$ is continuous on $[0, \infty) \times [0, \infty)$ and takes value zero at $(0, \delta_P), (\delta_A \to \infty, \delta_P)$. By envelope theorem, the effect on principal's expected payoff is

$$\frac{\partial EU_P}{\partial \delta_P} = -\frac{1}{2}(a - a^*)^2 \tag{28}$$

Hence, the principal with $\delta_P = 0$ is better off than any principal with $\delta_P > 0$. ■

References

[1] Akerlof, G. A.; Kranton, R. E. (2005): Identity and the Economics of Organizations. The Journal of Economic Perspectives 19, 9-32.

[2] Biel-Rey, P. (2002): Inequity Aversion and Team Incentives. UFAE and IAE Working Papers 677.07.

[3] Bolton, P.; Dewatripont, M. (2005): Contract Theory. MIT Press.

[4] Bolton G.E., Ockenfelds A. (2000): ERC: A Theory of Equity, Reciprocity, and Competition. American Economic Review 90:1, 166-193.

[5] Charness, G.; Dufwenberg, M. (2006): Promises and Partnership. Econometrica.74, 1579-1601.

[6] Charness, G.; Rabin, M. (2002): Undestanding Social Preferences with Simple Tests. Quarterly Journal of Economics 117, 817-869.

[7] Dufwenberg, M. (2002): Marital Investment, Time Consistency & Emotions. Journal of Economic Behavior & Organization 48, 57-69.

[8] Dufwenberg, M.; Gneezy U. (2000): Measuring Beliefs in an Experimental Lost Wallet Game. Games & Economic Behavior 30 (2000), 163-82

[9] Dufwenberg M., Kirchsteiger G.(2002): A Theory of Sequential Reciprocity. IUI working paper. University of Stockholm.

[10] Dur, R.; Glazear, A. (2004): Optimal Incentive Contracts when Workers Envy Their Boss. Tinbergen Institute Discussion Paper 2004-046/1.

[11] Engelmaier F, Wambach A (2005): Optimal Incentive Contracts under Inequity Aversion. IZA Discussion Paper Series, No. 1643.

[12] Engelmann, Dirk, and Martin Strobel (2004): Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments. American Economic Review 94, 857–69.

[13] Fehr, Ernst and Schmidt, Klaus M., (1999). A Theory of Fairness, Competition and Co-operation. Quarterly Journal of Economics 114, 817-868.

[14] Güth, W., and M.E. Yaari (1992): Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach, in Explaining Process and Change: Approaches to Evolutionary Economics, (ed.) U. Witt, University of Michigan, Ann Arbor.

[15] Holmström, B. (1979): Moral Hazard and Observability. Bell Journal of Economics 10, 74-91.

[16] Holmström B, P. Milgrom (1987): Aggregation and Linearity in the provision of Intertemporal Incentives. Econometrica 55, 303-328.

[17] Huck S., D. Kübler,J. Weibull.(2006): "Social Norms and Economic Incentives in Firms", Else Working Paper. University College London.

[18] Itoh, H. (2004): Moral Hazard and Other-Regarding Preferences. Japanese Economic Review 55, 18–45.

[19] Jewitt (1988): Justifying the First-Order Approach to Principal-Agent Problems. Econometrica 56, 1177-1190

[20] Miettinen, T. (2006): Promises and Conventions - An Approach to Pre-play Agreements. Max Planck Institute Discussion Paper on Strategic Interaction.

[21] Millar, K.U., Tesser A. (1988): Deceptive Behavior in Social Relationships: a Consequence of Violated Expectations. Journal of Psychology 122, 263-273.

[22] Rabin, Matthew (1993): Incorporating Fairness into Game Theory and Economics. American Economic Review 83, 1281-1302.

[23] Rob, R.; Zemsky, P. (2002): Social Capital, Corporate Culture, and Incentive Intensity. Rand Journal of Economics 33, 243-257.

[24] Samuelson, L. (2001): Introduction to the Evolution of Preferences.Journal of Economic Theory 97: 225-230

[25] Sobel, J. (2005): Interdependent Preferences and Reciprocity. Journal of Economic Literature 43, 392–436.