Moral Hazard and Clear Conscience[*]

Topi Miettinen
Max Planck Institute of Economics[†]
RUESG, University of Helsinki

July 9, 2007

Abstract

The hidden-action moral hazard model is one of the cornerstones of contract theory and it has become a basic tool to analyze job contracts and other agency relationships. Close social ties often implied by such relationships make the moral hazard model a prime application of approaches where agents care about inequity, and intrinsically reciprocate each others' kind and hostile intentions.

We contribute to this literature by considering how the optimal contract is affected if agreed with an agent who feels bad when not reaching a target effort set in the contract. Guilt makes effort partially contractible even without any monitoring. Not surprisingly, higher effort can be implemented with lower risk and the solution is closer to first-best.

Nevertheless, using the target effort to induce effort is not entirely costless for the principal. In equilibrium, the agent fails to provide the required effort. This induces guilt which must be compensated for by the principal, for otherwise the agent would not accept the job in the first place. Thus, both the principal and the agent are more pleased with the generated monetary return unlike suggested by previous social preference applications to optimal contracts.

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

[*] Thanks to Martin Dufwenberg, Florian Englmaier, Steffen Huck, Mikko Leppämäki, Eva-Maria Steiger, Hannu Vartiainen, and Torsten Weiland for comments and Yrjö Jahnsson Foundation for financial support.

[†] 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. With a flat wage scheme, there is no reason to provide effort. The quotation 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 the 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 'no': 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 hidden action 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 earns a better pay in terms of the certainty equivalent, even if her monetary incentives to provide effort are lower[1].

[1] Thus an evolutionary model could be put forth to argue that proneness to guilt has an evolutionary explanation (Güth and Yaari, 1992; Samuelson, 2001).

Formally, we introduce two additional features into the model of 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, otherwise it would be mere cheap talk with no impact. 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. Since guilt makes talk costly, any target that differs from the agent's optimal choice may have an impact on effort.

Guilt makes effort partially contractible even without monitoring. Yet surprisingly, using the target effort as an incentive mechanism is not entirely costless for the principal. 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, Huckleberry's employer gets higher earnings than if Huckleberry felt equally good about doing right and doing wrong and, moreover, Huckleberry is more pleased with his monetary remuneration even taking into account the higher effort he exerts.

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 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.

A bulk of literature considers the effects of agents' equity concerns on optimal contracts. The models closest to our setup are those of Englmaier and Wambach (2005), Itoh (2004), and Dur and Glazer (2004) who consider an agent who envies his principal[2]. Unlike the current paper, all these models assume risk-neutral parties and/or contractible effort. Thus they do not make behavioral departures from the general model of Holmström (1979), but rather from simpler setups. They all find that the principal's equilibrium payoff decreases and that of the agent increases if the agent is more concerned about equity:

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

When failing to produce output, the agent can be paid according to her outside option compensation. Yet, 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 benefits both the agent and the principal.

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 of Akerlof and Kranton (2005), however, completely abstracts from the endogenous cost of inducing a higher target effort which is present in our model. 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 even when using the informal target is not directly costly.

The paper is organized as follows. Section 2 relates the guilt-aversion model to other social preference approaches. Section 3 presents the general moral hazard model with agent having a preference for clear conscience. An example provides some further intuition and results. Section 4 considers a principal with proneness to guilt. Section 5 concludes.

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 intrinsically value behavior that conforms with social ideals, 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 resentment and reduces payoff. This idea, although neglected for long, has recently gained a lot of interest in the realm of economics. The literature on social preferences[3] provides with numerous examples: intention-based reciprocity-models, (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004); outcome-based inequity models, (Fehr and Schmidt , 1999; Bolton and Ockenfels, 2000); and hybrid models (Falk and Fischbacher, 2006). Not explicitly incorporating social ideals into the players' preferences implies that a descriptive theory may fail to capture some essential features which may hinder thorough understanding, or even sharper predictions and implementation of better economic policies.

Yet, it is not evident that one of the fairness or the reciprocity approaches provides a simple and unified explanation 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 are internalized and that can be formalized as a social utility function by theorists. The concern for equity is just one particular principle of distributive justice. Similarly, reciprocity could be thought of as a concern to contribute to the 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 will provide the ultimate salient explanation[4].

That a social ideal is common knowledge seems like a good way to formalize saliency in game theoretic terms. 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

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

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

potentially complex interaction context.

However, the first problem does not exist and the second is alleviated if the socially ideal action itself is 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 first inferred from the context.

How are such commonly known action-norms established? Clearly, face-to-face preplay negotiations, the specific focus of our paper, provides a means to achieve that. 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 violating the norm and, more importantly, that the guilt is increasing in the harm that the violation 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 from deviating.

3 Agent with a Preference for Clear Conscience

3.1 The general case

Let us consider Holmström's (1979) single-dimensional moral hazard model. The risk-neutral principal owns a stochastic production technology. The output level is denoted by q with support $[\underline{q}, \overline{q}]$. The 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

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

where $f(q; a)$ is the density function of the output q. The 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

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

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 $F_a(q;a'')$.

The agent's utility is additively separable in money and effort. A von Neumann-Morgernstern utility function $u(s(q))$ captures the agent's risk preferences over wage lotteries. The agent is strictly risk-averse

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

The agent suffers an effort cost captured by function $c(a) : R_+ \longrightarrow R_+$ which is increasing and convex on the set of actions and $c(0) = 0$

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

We add a behavioral component to the agent's preferences: the agent has preference for clear conscience meaning that she suffers from guilt if she inflicts harm on the principal by providing less effort than agreed. This component is additively separable from the other two. For simplicity, we assume a specific form of the clear conscience utility function where $g = \frac{1}{2}(\min\{a - a^*, 0\})^2$ so that implicitly the agent suffers only if she harms the principal by violating the agreed target effort denoted by a^*. 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

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

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 to the agent. This proposal consists of two parts: a monetary incentive scheme, $s(q)$, and a target action, a^*. The 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 a guilty conscience. The agent has an outside option $\underline{u}$ which captures the opportunity cost of the agent.

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

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

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

$$s.t.$$

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

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

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. The effort level chosen by the agent is below the target effort level. The 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

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

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

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

[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.

The latter is a result analogous to that in Holmström (1979) and it states that the monetary reward is increasing in the 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 the 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, suffers from a 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 a higher certainty equivalent than a zero proneness counterpart with equal risk attitudes 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]. 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 and for exerting a higher effort to make her accept the job in the first place. Hence, the gain from the monetary remuneration net of the cost of effort will be above the outside option payoff whereas for the zero proneness to guilt agent, this equals the outside option payoff.

Yet, it does not pay off for the agent to have an infinite proneness to follow the contract, because then the contract requires the agent to provide first-best effort and the agent obeys fully. The certainty equivalent net of the physical

[8] Consider, a job where effort is perfectly monitored, for instance.

cost of effort now coincides with the outside option payoff.

Proposition 2 *The certainty equivalent (gross/net of physical cost of effort) for an agent with proneness to guilt $\delta \in (0, \infty)$ is higher than that of an agent with zero proneness to guilt. The certainty equivalent (net of physical cost of effort) of an agent with proneness to guilt $\delta \in (0, \infty)$ is also higher than that of an agent with infinite proneness to guilt.*

Proof. It is easy to see that, in equilibrium, the agent's gain from monetary remuneration net of the physical cost of effort satisfies

$$\begin{aligned}\int u(s_\delta(q))f(q;a_\delta)dq - c(a_\delta) &> \int (s_\delta(q))f(q;a_\delta)dq \\ &\quad -\frac{\delta}{2}(a_\delta - a^*)^2 - c(a_\delta) \\ &= \underline{u} \\ &= \int u(s(q))f(q;a)dq - c(a_0),\end{aligned}$$

where a_δ and a_0 are the equilibrium effort levels chosen by the agent with proneness δ to clear conscience and zero proneness to guilt respectively. Also, $c(a_\delta) > c(a)$ since $a_\delta > a$. Hence, an agent with positive δ has a higher certainty equivalent than an agent with the same risk attitude and with $\delta = 0$. Moreover also the gain from monetary remuneration net of the physical cost of effort is greater.

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 the principal faces an agent with $\delta > 0$ than when the 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 utility gained by an agent with $\delta > 0$ is at least weakly higher than that of an agent with $\delta = 0$ whether the emotional term is included in the utility or not. Thus, 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. This is because the target effort reduces the cost of implementing inframarginal units of effort due to partially avoiding the incentive pay which generates risks to the agent. Optimality, with a risk neutral principal, requires that the expected marginal productivity of effort equals the marginal cost of implementing it. The equilibrium effort level and equilibrium productivity will thus be higher than in a model without proneness to guilt, which further increases overall welfare.

Providing 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 the realized action is not observable or enforceable. Existence of clear conscience preferences may be one explanation.

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. Moreover, the agent's utility components are multiplicatively separable. In this simple case, the optimal solution can be explicitly derived 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 $q\tilde{}N(a, \sigma^2)$

(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 $u(y) = -\exp(-ry)$.

[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.

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 who is 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 does not have an explicit preference 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.*

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

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:

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

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. Thus the difference between the certainty equivalents net of physical cost of effort between 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

$$\Pi(\delta) = \frac{\delta(r\sigma^2)^2}{2(1+(c+\delta)r\sigma^2)^2} > 0.$$

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

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

whereas clearly for $\delta > 0$, $\Pi(\delta) > 0$.

Remark 8 *In terms of the certainty equivalent net of physical cost of effort, the agent who is infinitely prone to guilt is equally well of as an agent with zero proneness to guilt.*

There is a unique value,

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

which maximizes the certainty equivalent net of physical cost of effort.

4 Principal with Preference for Clear Conscience

In this section, we consider a scenario where 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.

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 9 *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 $\delta_P \in (0, \infty)$.*

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 a 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 certainty equivalent gross/net of physical cost of effort than an agent with zero proneness to guilt.

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.

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.; Glazer, A. (2004): Optimal Incentive Contracts when Workers Envy Their Boss. Tinbergen Institute Discussion Paper 2004-046/1.

[11] Englmaier 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.