Moral Hazard and Clear Conscience*

Topi Miettinen
SITE, Stockholm School of Economics†

July 21, 2009

Abstract

The paper studies theoretically how the optimal contract in the hidden-action moral hazard is affected when an agent feels bad when not reaching a target effort set in the contract. Agent's guilt or grief about not reaching the target 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 to induce effort is not entirely costless for the principal. In equilibrium, the agent's effort falls short of the target. 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, although the principal's payoff is higher, the agent receives a part of the monetary rents accruing to intrinsic motivation. This result differs markedly from those in previous contributions on contracting with an agent with social preference or pro-social motivation.

KEYWORDS: Moral Hazard, Norms, Agency, Social Preferences, Guilt, Work Ethic

JEL: C72, D82, Z13

*Thanks to Martin Dufwenberg, Robert Dur, Florian Englmaier, Steffen Huck, Marieke Huysentruyt, Mikko Leppämäki, Eva-Maria Steiger, Ute Stephan, Hannu Vartiainen, Torsten Weiland, and two anonymous referees for comments. Financial support of Yrjö Jahnsson Foundation and the European Union within the project "Social Entrepreneurs as Lead Users for Service Innovation" (grant agreement 217622). Part of the research was carried out when the author was affiliated to Max Planck Institute of Economics, Jena, Germany.

†Sveavägen 65, PO Box 6501 SE-113 83 Stockholm Sweden; topi.miettinen@hhs.se

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, or the socially desirable, 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 in pecuniary terms from feeling better about doing the 'right' or promoting the good of society? 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.

Building on the single dimensional hidden action moral hazard model (Holmström, 1979), I illustrate 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. I 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.

Formally, I 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 mere cheap talk with no impact and thus left unmodelled. The second new and a very much related ingredient is that the agent may have a preference for clear conscience: she suffers a cost if she fails to meet the target. Since guilt makes talk costly, the target may have a direct impact on effort.

The target effort might refer to an informal non-binding mutual agreement between the principal and the agent or a goal assigned to the worker. The positive impact of promises on efficiency and on the principal's payoffs has been recently empirically documented by Charness and Dufwenberg (2006) and Vanberg (2008), and modeled by the former as well as Dufwenberg and Battigalli (2007, 2009) and Miettinen (2006). The positive impact on effort of non-binding non-incentivized goals has been widely documented and theorized by psychologists (Locke and Latham, 2002; Latham and Locke, 2007).

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 Huckleberry is willing to accept the contract. However, since the adopted incentive scheme is less risky than one which does not take advantage of Huckleberry's moral preferences, Huckleberry's employer gets higher earnings than if Huckleberry felt equally good about doing right and doing wrong and, moreover, Huckleberry earns a higher certainty equivalent, even net of the higher effort he exerts. The findings are aligned with those of experimental contract theory literature. Fehr et al. (1997, p.849) and Fehr and Rockenbach (2003), for instance, find that an optimally chosen combination of higher target effort and wages (above opportunistic subgame-perfect equilibrium levels) benefits both the principal and the agent, although actual effort falls short of the target effort. Reciprocity provides an alternative explanation for the findings.[1]

Section 3 considers the case where, in addition to the agent, also the principal is motivated by preference for doing right. Once the output has been produced, it is in the interest of the principal not to pay the agent but rather to keep the entire value of the output to herself. I illustrate that a principal with observable

[1]There is thus room for experiments that attempt to disentangle the reciprocity and guilt effects in these settings. Dufwenberg and Kirchsteiger (2004) and Falk and Fishbacher (2006) provide general formalizations of reciprocity.

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. As in Saloner and Rotemberg (1993), other-regarding preference provides intrinsic commitment power to the principal which, in a context where contracts are highly incomplete, is necessary for credible provision of incentives to the agent.

A bulk of literature considers the effects of agent's equity concerns on optimal contracts. The models closest to my setup are those of Englmaier and Wambach (2005), Itoh (2004), and Dur and Glazer (2008) who consider an agent who envies his principal. Unlike the current paper, all these models assume risk-neutral parties and/or contractible effort. Thus they build upon simpler or different setups than the model of Holmström (1979). They all find that the principal's equilibrium payoff decreases and that of the agent increases if the agent is more concerned about equity: 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 the present one 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 my model since the identity is essentially a preference for doing as the identity calls for.[2] 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, abstracts from the endogenous cost of inducing a higher target effort which is present in my model. Rather, they assume an exogenous cost of building up corporate identity. Apart from the exogenous costs/benefits of corporate 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 presents the general moral hazard model with agent having a preference for clear conscience. An example provides some further intuition and results. Section 3 considers a principal with proneness to guilt. Section 4 concludes.

[2]Alternatively, the target effort might reflect the agent's work ethic or a social norm which again may be manipulated by the principal.

2 Agent with a preference for clear conscience

2.1 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 effort $a \in [0, \infty)$ and output is drawn randomly from a distribution that is parametrized with effort. The cumulative distribution function is $F(q; a)$. Let's 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 $F_a(q; a'')$ - greater effort is more likely to result in higher output.

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 a disutility of effort captured by function $c : R_+ \longrightarrow R_+$ which is increasing and convex in effort and $c(0) = 0$,

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

I add a behavioral component to the agent's utility function: the agent suffers disutility of 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, I assume a specific form of this disutility denoted by g such that $g : R \longrightarrow R_+$ $g(a, a^*) = \frac{1}{2}(\min\{a - a^*, 0\})^2$. Thus the agent suffers only if she harms the principal by falling short of the agreed target effort denoted by a^*. Notice that the disutility of guilt is increasing in the harm inflicted on the principal.[3] In summary the function of disutility of effort of an agent with

[3]The truncation is made for plausibility. The results would be unchanged if rather $g =$

proneness to guilt δ 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. I assume that the agent's preferences and costs are observable to the principal. [4]

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 effort, 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 or feels devastated by her lacking work morale. 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[5]

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

$$s.t.$$

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

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

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.*

$\frac{1}{2}(a - a^*)^2$ was assumed as is clear from the proof of Proposition 1, equations (9) and (10) in particular. For the sake of simplicity, I do not make guilt a function of the expected harm of the principal explicitly.

[4] Alternative assumptions are discussed in the conclusion.

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

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

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

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. Kuhn-Tucker conditions imply that λ is 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

$$\int u(\frac{1}{u'} - \lambda)f(q;a)dq = \mu(c'(a) + \delta(a - a^*)) \tag{9}$$

Taking the expectation on both sides of (7) gives

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

Then the left hand side of (9)

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

has the sign of covariance of $\frac{1}{u'}$ and u. Since u and $\frac{1}{u'}$ are monotone in the same direction, this sign is positive. Therefore, on the left hand side of (9) 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

$$sgn(\mu) = sgn(c'(a) - \delta\frac{\mu}{\lambda}) \tag{10}$$

Hence μ cannot be non-positive because with non-positive μ, the right hand side of (10) is strictly positive and the equality does not hold. Hence, μ must be strictly positive. This implies that

$$a^* > a.$$

Setting $a^* \leq a$ cannot be optimal since then guilt would be zero and the constrained optimum for $a^* \leq a$ is the contract for the standard agent which was shown above to be suboptimal if $\delta > 0$.

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. One can consider the principal's strategy of setting the target effort above the equilibrium effort as a means of generating intrinsic marginal gains from higher effort for the agent - those of reducing disutility of guilt. The guilt prone agent equates the marginal physical disutility of effort with the marginal gains, which in this case consist of the expected marginal increases in monetary remuneration and, in addition, of the reduction in the marginal disutility of guilt. The higher above the actual effort the target lies, the greater the marginal intrinsic gain for the agent from increasing effort. High-powered monetary incentives and target effort are thus imperfect substitutes in inducing effort where the substitutability depends on the agents proneness to guilt. Both the high powered monetary incentives and high targets are detrimental for the agent because strong pecuniary incentives make income risky and because a higher target increases the total disutility of guilt, respectively. In the optimal contract, the principal trades off the gains and costs of using these alternative means of eliciting effort.

The following corollary 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 proneness-to-guilt types have equal wages in an outside option. In equilibrium the agent prone to guilt suffers since she never reaches the target effort. The agent must be compensated for having to feel guilt and for exerting higher effort to make her accept the job in the first place. Hence, the gain from the monetary remuneration net of the physical disutility of effort will be above the outside option payoff whereas for the zero proneness to guilt agent, this equals the outside option payoff.

Corollary 1 *The certainty equivalent (gross/net of physical disutility 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.*

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

$$\begin{aligned} \int u(s_\delta(q))f(q;a_\delta)dq - c(a_\delta) &> \int u(s_\delta(q))f(q;a_\delta)dq \\ &\quad -\frac{\delta}{2}(a_\delta - a_\delta^*)^2 - c(a_\delta) \\ &= \underline{u} \\ &= \int u(s_0(q))f(q;a_0)dq - c(a_0), \end{aligned}$$

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, $c(a_\delta) > c(a_0)$ since $a_\delta > a_0$. 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 disutility of effort is greater.

■

Notice that principal could induce the effort choice of the standard agent from any type of agent such that even the costs and benefits to both parties of

the agreement would be invariant to the proneness to guilt type of the agent. This would require using, for any proneness to guilt type, the same monetary remuneration scheme as for the standard agent and setting the target effort equal to the equilibrium effort of the standard agent. By choosing the equilibrium effort of the standard agent, the agent prone to guilt would thus suffer from no disutility of guilt and even the expected utility apart from this term would be maximized. Any deviation would therefore be detrimental. Since the principal prefers using another contract when the agent is disposed to feel grief about not reaching targets, the principal must gain from contracting with an agent whose proneness to guilt is positive. Given Corollary 1, it is thus straightforward that efficiency is higher as well.

Corollary 2 *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. This contract would be accepted by every type of agent since the optimal effort of each agent would then coincide with that of the agent with $\delta = 0$ - agents with $\delta > 0$ suffer even more than agents with $\delta = 0$ from deviations from this optimum due to their conscience preference. Thus each δ yields the same expected payoff and suffers no guilt. The participation constraint is satisfied since it is for $\delta = 0$.

It is shown above that proposing the optimal contract for $\delta = 0$ to an agent with $\delta > 0$ is suboptimal and yet the contract for $\delta = 0$ performs equally well for $\delta = 0$ and for any $\delta > 0$. Thus, the principal must strictly gain from contracting with $\delta > 0$ rather than $\delta = 0$. It is also shown that the utility of 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 lessening the need for incentive pay which generates risks to the risk-averse 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 when the agent is not disposed to feel bad about not reaching targets. This 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 effort is not observable or enforceable. Disposition to feel grief about not reaching targets may be one explanation. [6]

Qualitatively, the results would continue to hold even if the agent was risk-neutral and the contract was made subject to limited (zero) liability of the agent.[7] First, if the agent could walk out of the contract and receive her outside-option payoff at any time (zero liability) without having to feel guilty, then even the limited liability constraint implies that the agent will be remunerated for having to bear guilt since otherwise the agent would indeed walk out of the contract when unsuccessful. It is of course somewhat implausible that the agent could choose not to feel guilty at all by exiting. Yet, if there is guilt about quitting ex post then it is not the liability constraint but the ex-ante participation constraint that implies the remuneration for equilibrium guilt.

2.2 Linear-normal example

By means of an example we shall consider here 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 effort. Moreover, the agent's utility components are multiplicatively separable. In this simple case, the optimal solution can be explicitly derived and thus it can used to better illustrate the intuition of the model[8]. The assumptions of the model are as follows:

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

[6] There are other explanations as well, such as the principal avoiding to divert the agent's attention away from other activities important for profitability when the agent has several tasks to carry out at the same time (Holmström and Milgrom, 1991).

[7] See Sappington (1983), for instance.

[8] 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 physical disutility 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 $y = vq+t-\frac{c}{2}a^2-\frac{\delta}{2}(a-a^*)^2$.

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 second constraint can be alternatively written as

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

Solving the system is straightforward and thus proofs are dropped. The optimal contract reads

$$v_\delta = \frac{1}{(1 + (c + \delta)r\sigma^2)}, \tag{12}$$

$$a^*_\delta = \frac{1}{c}. \tag{13}$$

$$t_\delta = \underline{u} + \frac{(rc\sigma^2 - 1) + \delta(c + \delta)(r\sigma^2)^2}{2c(1 + (c + \delta)r\sigma^2)^2}. \tag{14}$$

From (11), it is now easy to see that the monetary incentives and the target effort are imperfect substitutes in inducing effort. Moreover, an agent who is more prone to guilt is less responsive to monetary incentives than a standard agent since the former, rather than equating the marginal disutility of effort to the marginal expected monetary gain, also takes into account the diminishing marginal disutility of falling behind the target as effort increases. Thus, while at optimum the guilt-prone agent typically exerts more effort than a standard

agent, the diminishing marginal disutility of guilt implies that the agent is less responsive to monetary incentives.

Independently of the agent's proneness to guilt, the principal sets the target effort equal to the first best effort of the agent[9]. The agent prone to guilt is offered a lower bonus rate than the agent with zero proneness to guilt.

The target effort is always above the equilibrium effort:

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

and fixed remuneration for the guilt prone agent is higher and, yet, the guilt-prone agent chooses a higher effort level than the agent without proneness to guilt. The risk neutral principal bears a larger share of the risk, which further improves efficiency.

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:

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

Notice that $v_\delta a_\delta + t_\delta - \frac{r v_\delta^2 \sigma^2}{2} - \frac{c}{2}a_\delta^2$ is the agent's material payoff. Thus the difference between the certainty equivalents net of physical disutility of effort between a guilt-prone agent and a standard agent is merely $\Pi(\delta) \equiv \frac{\delta}{2}(a_\delta - a_\delta^*)^2$. Plugging (15) into $\frac{\delta}{2}(a_\delta - a_\delta^*)^2$ we obtain the difference

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

Applying Hospital's rule gives

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

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

One can also make the remark that it does not pay off for the agent to have too large a proneness to guilt. If the disposition to disutility of guilt is too important relative to that to the physical disutility of effort, the agent will not miss the target by a large margin. Thus disutility of guilt will ultimately be vanishing in the agent's proneness to guilt and so will be her share of the

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

monetary rents generated by this disposition. There is in fact a unique value,

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

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

3 Principal with preference for clear conscience

In this section, we shall 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 produced 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.

There are three stages in the game. First, the principal and the agent agree on an incentive scheme $s^*(q)$. In the second stage given the principal's proneness to guilt and the incentive scheme $s^*(q)$, the agent chooses effort a and the stochastic output q is produced according to $F(q;a)$.[10] In the third stage, the principal pays out the actual remuneration $s(q)$ which may be conditioned on the output.

The principal's disposition to grief is modelled in the following manner. The principal is assumed to suffer from guilt if her ex-post payment to the agent, $s(q)$, is smaller than the amount indicated in the agreed-upon incentive scheme, $s^*(q)$. The disutility of guilt is increasing in the harm inflicted on the agent such that $g : R \longrightarrow R_+$ and $g(s, s^*) = \frac{1}{2}(\min\{s - s^*, 0\})^2$ where s is the actual payment paid out to the agent when output is produced and s^* is the payment promised to the agent when contracting. Thus, given q, the principal's ex-post problem amounts to the maximization of

$$\max_s U_P(q, s, s^*(q); \delta) =$$
$$\max_s \{\pi_P(q, s) - \frac{\delta_P}{2}(min\{0, s - s^*(q)\})^2\}$$

[10] For simplicity, we assume that the agent is a standard one throughout this section.

where $\pi_P(q,s) = q - s$ is the material payoff and $\frac{\delta_P}{2}(min\{0, s - s^*(q)\})^2$ is the principal's disutility of guilt. If $\delta_P \in (0,\infty)$, the problem admits an interior solution

$$s(q) = s^*(q) - \frac{1}{\delta_P} \tag{16}$$

whereas the principal with no disposition to bad feelings about breaching chooses $s(q) = 0$ for all q.

Let's now turn attention to the agent's optimal effort. Given that the disposition characteristics of the principal are observable, the agent perfectly anticipates the lack of commitment of the principal and foresees the bias given in (16). Thus, the principal with $\delta_P \in (0,\infty)$ can implement the scheme of a fully committed principal $\hat{s}(q)$ by setting $s^*_\delta(q) \equiv \hat{s}(q) + \frac{1}{\delta_P}$. Let us denote the agent's optimal effort choice under the scheme $\hat{s}(q)$ by $\hat{a}$. Indeed, notice that if dispositional commitment power approaches full commitment power $\delta_P \to \infty$, then $s^*(q) \to \hat{s}(q)$. It is easy to see that for every $\delta_P \in (0,\infty)$ this constitutes the ex-ante optimal scheme if the principal decides to contract at all. It is also easy to see that all principals with $\delta_P \in (0,\infty)$ receive exactly the same expected monetary payoff $\overline{\pi}_P \equiv \int_{\underline{q}}^{\overline{q}} (q - \hat{s}(q)) f(q;\hat{a}) dq$.

Consider now the contracting phase. To fix ideas, suppose that the principal has a disposition-independent alternative use for the assets she owns with expected return $\underline{\pi}_P$, such as selling the assets or controlling the production technology oneself. The equilibrium disutility of guilt is, perhaps surprisingly, a decreasing function of the principal's proneness to guilt and equals $\frac{1}{\delta_P}$. Thus the disposition to guilt may influence the principal's optimal choice between hiring an agent or allocating the assets to the alternative use. To see this, denote by $\underline{\delta}_P$ the proneness to guilt type of the principal who is indifferent between offering the contract to the agent and choosing the outside option. This type is implicitly given by $\int_{\underline{q}}^{\overline{q}} (q - \hat{s}(q)) f(q;\hat{a}) dq - \frac{1}{\underline{\delta}_P} = \underline{\pi}_P$. Solving for $\underline{\delta}_P$ yields

$$\underline{\delta}_P = \frac{1}{\int_{\underline{q}}^{\overline{q}} (q - \hat{s}(q)) f(q;\hat{a}) dq - \underline{\pi}_P}.$$

A principal with $\delta_P < \underline{\delta}_P$ will refrain from hiring an agent altogether since her expected utility is below her expected return in the outside option.

Finally, any credible incentive scheme 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 choose her outside option $\underline{u}$.

Proposition 2 *The expected monetary return for the principal with* $\delta_P \leq \underline{\delta}_P$ *is* $\underline{\pi}_P$ *and the expected monetary return for the principal with* $\delta_P > \underline{\delta}_P$ *is* $\overline{\pi}_P$ *where* $\overline{\pi}_P > \underline{\pi}_P$*. The expected utility of the principal with* $\delta_P \leq \underline{\delta}_P$ *equals* $\underline{\pi}_P$*. The utility of the principal is increasing in* δ_P *in* $[\underline{\delta}_P, \infty)$ *and* $\lim_{\delta \to \infty} \int_{\underline{q}}^{\overline{q}} U_P(q, \hat{s}(q), s^*_\delta(q); \delta) f(q; \hat{a}) dq = \overline{\pi}_P$.

4 Discussion

Recent experimental evidence from controlled laboratory experiments indicates that people tend to avoid breaching informal promises (Charness and Dufwenberg, 2006; Vanberg, 2008). Psychologists in turn have shown that specific non-binding targets increase performance more than just urging people to do their best (Locke and Latham, 2002). These findings may accrue to the tendency of feeling guilt or grief about not keeping promises and meeting targets. In this paper, I have theoretically studied the effects of preference for clear conscience on equilibrium contracts and behavior in a hidden action moral hazard setup.

As for the results, when facing *an agent prone to guilt*, the principal finds it profitable to set the target effort above the equilibrium effort of the agent and to use equilibrium guilt to better incentivize the agent. Because the agent must be compensated sufficiently for the grief to accept the job, an agent prone to guilt receives a higher certainty equivalent gross/net of physical disutility of effort than an agent with zero proneness to guilt. On the other hand, 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 nothing will not accept the contract.[11]

The results can also be approached from the angle of contracting under pro-social motivation (Francois and Vlassopoulos, 2007). The target might reflect the socially optimal action, especially if production generates unmodeled positive externalities on third parties. The principal might then be in a position of manipulating social cues that easily trigger the agent's pro-social motivation to work harder. Pro-social cuing of a task with positive externalities on third

[11] Bester and Strauss (2001) analyze such commitment problems in a setting where the initial contract offer restricts the set of actual payments.

parties strengthens latent pro-social identity so that the socially optimal effort becomes the focal target. A row of papers have established results on improved risk-sharing, lower-powered incentives, and lower total monetary compensation for pro-socially motivated agents (see Francois, 2001, for instance). The present paper suggests that if guilt motivation is important among pro-socially motivated agents, then all monetary rents from pro-social motivation may not accrue to the principal but the agent yields a higher compensation than in tasks where pro-social cuing is infeasible.

In this paper the focus has been limited to the implications of clear conscience preferences on optimal contracts. This class of preferences is empirically well-documented in controlled economic experiments. Yet, an interesting extension of the present model would allow agents being proud of exceeding their targets, in addition to feeling bad about not reaching them. In accordance with other behavioral domains (Tversky and Kahneman, 1979; Fehr and Schmidt, 1999), where people have been found to perceive losses more salient than gains, the marginal pride in the gain domain would perhaps be best modeled smaller than the marginal guilt in the loss domain. If both gains and losses are concave, the principal would face a trade-off between easier targets which the agent would surpass and more challenging targets that the agent would never reach. The former would be less expensive for the principal but would also imply weaker intrinsic incentives whereas the latter provide strong incentives, yet with a cost to the principal. Thus, depending on the production technology, the former or the latter class of targets may be optimal.

Despite the loss in parsimony and the lack of rigorous experimental economics research on pride, the extension to the gains domain suggests itself in psychological goal-setting theory (Latham and Locke, 2007) where satisfaction over beating the targets plays an important role. Psychologists argue that persistent falling short of targets reduces empathy towards the employer over time. Thus guilt based incentives are likely to be of limited applicability in dynamic contexts. Anecdotal evidence from investment banking and consultancy agencies suggests that targets superior to equilibrium effort have been used, especially for those in early phase of their career path. Due to the interest of screening out and inducing the exit of worst performers when agent's type is private information, targets with equilibrium guilt may be a part of an optimal strategy after all provided that some types can reach the targets. In contrast, satisfaction about beating targets is argued to improve empathy over time and thus easy targets provide the added value of aligning the agent's incentives with

those of the organization in future interaction.[12]

In this paper, for simplicity, it is assumed that the agent's proneness to guilt and his skill is fully observable. Yet, in the light of Frank (1987), the results could be generalized to allow for only partial observability of clear conscience preferences. Further research is certainly needed in order to understand how non-pecuniary motivation, including the motivation suggested in this paper, twists incentives and influences optimal contracts.

References

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

[2] Battigalli, P., Dufwenberg, M., 2007. Guilt in Games. American Economic Review, P&P 97, 170-76.

[3] Battigalli, P., Dufwenberg, M., 2009. Dynamic Psychological Games. Journal of Economic Theory 144, 1-35.

[4] Bester, H.. Strauss, R., 2001. Contracting with Imperfect Commitment and the Revelation Principle: The Single Agent Case. Econometrica 69, 1077-1098.

[5] Bolton, P., Dewatripont, M., 2005. *Contract Theory*. MIT Press.

[6] Charness, G., Dufwenberg, M. 2006. Promises and Partnerships. Econometrica 74, 1579-1601.

[7] Dufwenberg, M., Kirchsteiger, G., 2004. Theory of Sequential Reciprocity. Games and Economic Behavior

[8] Falk, A., Fishbacher, U., 2006. A Theory of Reciprocity. Games and Economic Behavior

[9] Ellingsen, T.; Johanneson, M., 2008. Pride and Prejudice: The Human Side of Incentive Theory. American Economic Review, 98, 990-1008.

[10] Englmaier, F., Wambach, A., 2005. Optimal Incentive Contracts under Inequity Aversion. IZA Discussion Paper Series, No. 1643.

[12] A related static model, yet where targets play no role, is put forward by Ellingsen and Johanneson (2008) where agents have preferences over the type of the principal they work for.

[11] Fehr, E., Gächter, S., Kirchsteiger, G. 1997. Reciprocity as a Contract Enforcement Device, Experimental Evidence. Econometrica 65, 833-860.

[12] Fehr, E., Rockenbach, B. 2003. Detrimental Effects of Sanctions on Human Altruism. Nature 422, 137-140.

[13] Francois, P. 2001. Employee Care and the Role of Nonprofit Organizations. Journal of Institutional and Theoretical Economics 157, 443-464.

[14] Francois, P., Vlassopoulos, M. 2007. Pro-Social Motivation and the Delivery of Social Services. CESifo Economic Studies 54, 22-54.

[15] Frank, R., 1987. If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience? American Economic Review 77, 593-604.

[16] Holmström, B., 1979. Moral Hazard and Observability. Bell Journal of Economics 10, 74-91.

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

[18] Holmström B., Milgrom, P., 1991. Multitask principal-agent analysis: Incentive contracts, asset ownership, and job design. Journal of Law, Economics, and Organization 7, 24-52.

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

[20] Jewitt, I., 1988. Justifying the First-Order Approach to Principal-Agent Problems. Econometrica 56, 1177-1190.

[21] Latham, G., Locke, E.A. 2007. New Developments in and Directions for Goal-Setting Research. European Psychologist 12, 290-300.

[22] Locke, E.A., Latham G. 2002. Building a Practically Useful Theory of Goal Setting and Task Motivation. American Psychologist 57, 705-717.

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

[24] Saloner, G., Rotemberg, J., 1993. Leadership Style and Incentives. Management Science 11, 1299-1318.

[25] Sappington, D. 1983. Limited Liability Contracts between Principal and Agent. Journal of Economic Theory 29, 1-21.

[26] Vanberg, C., 2008. Why Do People Keep Their Promises? An Experimental Test of Two Explanations. Econometrica 76, 1467-1480.