UNIVERSITY
OF TECHNOLOGY
SYDNEY

UTS Business School –
Department of Economics
Dr Chau Chak Wing Building
14 – 28 Ultimo Road, Ultimo NSW 2007

T: +61 2 9514 7713
economics@uts.edu.au

Department of Economics
Working Paper Series

'A Generalized Principal-Agent Model With Lying Costs'

Isa Hafalir [1]
Gordon Menzies [2]

[1] The University of Technology Sydney
[2] The University of Technology Sydney

No. 04 / 2023 **August 2023**

A GENERALIZED PRINCIPAL-AGENT MODEL WITH LYING COSTS

ISA E. HAFALIR AND GORDON MENZIES*

Abstract. Lie-aversion and lying costs should be included in models because disclosing hidden information generalizes basic theory. The agent in our principal-agent model has an exogenous lying cost. If it is high enough, she can be offered a first-best contract. If not, a modified contract outperforms the classic contract. If her cost is private information, lying occurs in equilibrium. The generalized theory suggests that the widespread offering of incentive contracts may initiate a 'vicious circle,' if it communicates untrustworthiness and lowers lying costs. Furthermore, cultures of untruthfulness may contribute to economic decline, and appearing dishonest can be rent-enhancing.

Keywords: Incentive contracts, Moral Hazard, Lying Costs, Optimal Mechanisms

JEL Classification Numbers: D82, D86

1. Introduction

The principal-agent model with hidden action (moral hazard) has been a workhorse in Economic theory for nearly half a century. In the context of the theory of the managerial firm, Jensen and Meckling (1976) is the seminal paper introducing moral hazard problems. In the basic model, the principal employs an agent to undertake a task earning stochastic payoffs for the principal, but the effort by the agent is unobservable. Therefore, the principal writes an incentive contract for the agent, paying more for good outcomes than bad ones, compensating her for the resultant risk.

Date: August 22, 2023.

(*) We would like to thank the Economic Humanist group, Donald Hay, David Vines, Tom Simpson, and Luis Pontes de Vasconcelos, all at Oxford University, and colleagues at UTS.

Hafalir (corresponding author): isa.hafalir@uts.edu.au, Department of Economics, University of Technology Sydney, PO Box 123, Broadway NSW 2007, Australia; Menzies: gordon.menzies@uts.edu.au, Department of Economics, University of Technology Sydney, PO Box 123, Broadway NSW 2007, Australia.

Despite their theoretical simplicity, principal-agent models have enjoyed remarkable influence in the practical world of banking, finance, and management and have been used to justify bonuses for super-managers, particularly in banking. Jensen himself took on a key leadership role at the Harvard Business School in the 1980s, influencing (along with Michael Porter) a whole generation of global managers (Khurana, 2010).

Their message was well-timed for an era of economic reform and openness to new ideas. Proponents successfully argued that because managers' work was not easily observable, they would fail to pursue actions that would maximize the value of the firm. The widely-implemented solution consisted of incentive contracts – paying managers by company shares or share options, so that their incentives aligned with the owners. Under the resultant profit-maximizing management, both parties were thought to financially benefit in ways that were socially optimal (Khurana, 2010).

Despite the reasonable expectation of optimality, the managers trained over this time were heavily criticized for a lack of integrity in the early years of the 21st century, following the 2008 Global Financial Crisis and other banking misadventures. The criticisms applied over a wide geographical area, embracing the UK (Salz and Collins, 2013; Martin, 2016) and the USA (Hill and Painter, 2020; Fligstein and Roehrkasse, 2016). Jensen himself noted a lack of truthfulness and a tendency to treat integrity as a matter of mere cost-benefit analysis (Jensen, 2009).

The contribution of this paper is to posit a connection between the widespread offering of incentive contracts, and subsequent untrustworthy behavior, using the notion of trust responsiveness. Trust responsiveness is the phenomenon that when someone feels they are being trusted, they are more likely to become trustworthy (Bacharach et al., 2007; Guerra and Zizzo, 2004). Conversely, if the offering of incentive contracts to managers somehow communicated that they were not trusted, it would have explained their becoming less trustworthy.

The way this lack of trust was communicated was through a radical change in the implicit assumptions made about truth-telling. Prior to the 1980s revolution in management training instigated by Jensen and Porter, it was common for the prinipal-agent problem to be addressed with a culture of professionalism (Lydenberg, 2014). The culture relied upon a different representative agent to 'economic man', the so-called 'reasonable person' who is more pro-social and hence more truthful (Menzies et al., 2019). This representative agent allows for the possibility of unincentivized truth-telling, and therefore it was thought that something like the first-best solution was feasible.

However, as the landscape was transformed in the 1980s, the classic principal-agent model became the standard way of thinking. In that model truth-telling is ruled out by making an implicit assumption that trustworthy communication is impossible. Otherwise, the principal could ask the agent what their effort level was and potentially receive a truthful answer.

The resultant change in culture was geographically widespread (Salz and Collins, 2013). The complaint was that bankers became entirely concerned with short term profitability, presumably because this would revalue their shares and share options made available through an incentive contract. In the Australian context, Hogan (2018) claims that from the 1990s the flow of foreign professionals (who would have been trained in the global Jensen and Porter paradigm) into the major banks led to a noticeable decline in concern for non-monetary aspects of banking.

It is an open question how common unincentivized truth-telling is, but it is beyond doubt that it sometimes happens. As discussed in Erat and Gneezy (2012), people may not like lying even if incentives point to it; hence there seem to be varying levels of lying costs, and truthful answers can be obtained for people with high lying costs. Naturally, this is to be distinguished from refraining from lying simply because it is optimal intertemporally (Gneezy et al., 2018). In another important paper, Abeler et al. (2019) combine data from 90 experimental studies and propose "preference for being seen as honest" and "preference for being honest" being the main motivations for truth-telling.

To represent integrity or its absence, the modeling device we use in this paper is a setup where an agent is asked to declare her effort in a generalized principal-agent model. If her declared effort and actual effort are different, the agent incurs a lying cost. This is most naturally thought of as part of their preferences, with a preference for being honest being represented by a high lying cost.

Our argument proceeds in two steps. First, we show what should be obvious: in a generalized principal-agent model with lying costs, the classic solution is a special case when the lying cost is zero. This implies that anyone offered a Jensen and Meckling (1976) contract is also receiving a 'message' that they are completely untrustworthy, in the sense that they are completely untroubled by lying.

In our simple setup, there is a risk-neutral principal and a risk-averse agent. The agent chooses either low effort or high effort, where high effort is more costly to the agent. There are two outcomes that bring either a high outcome or a low outcome, and a high (low) outcome is more likely following a high (low) effort. The principal also asks the agent whether she puts in a low or high effort, and the agent's actual effort need not align with their 'declared' effort. If the actual effort and declared effort differ, the agent pays a lying cost. The principal rewards the agent based on both the realized outcome and declaration.

We first analyze the case of 'known lying cost' and then extend our analysis to the private information lying cost case. Note that when lying cost is private information, we have both moral hazard (hidden action of what level of effort was chosen) and adverse selection (hidden information regarding the cost of lying.) To make our model simpler and to avoid extreme implausibility for any realistic application, the principal does not ask for the lying cost of the agent. Instead, the agent's reward is just based on declaration and outcome level, as in the known lying cost case. In the case of private information, some types of agents would be truthful, and some would lie.

If the lying cost is high enough, the agent reliably tells the truth, leading to the first-best solution (Proposition 2.) If the lying cost is low enough to disrupt the first-best solution, but not zero, the optimal incentive contract motivates the agent to both put in the high effort, and tell the truth about the effort level (Proposition 1.) If the lying cost is zero, we obtain the classic solution (Proposition 1) with its attendant 'message' that the agent is completely untrustworthy. But our analysis also shows that less extreme incentive contracts also contain a message about lying costs. As soon as the agent's wage starts to depend on outcomes rather than declarations, there is an unavoidable message of untrustworthiness sent. Depending on the assumed lying cost, this will vary from a message of 'completely' untrustworthy (the classic solution) to 'somewhat' untrustworthy for low-but-not-zero lying cost contracts. We also extend our analysis to the case where the lying cost is private information and obtain similar–but not identical–results (Propositions 3 and 4.)

In the second step of our argument, we model the impact of these messages of perceived untrustworthiness by reductions in agents' lying costs, setting in train a 'vicious circle.' (Proposition 5.) More specifically, for a given incentive structure, we assume that when agents first 'notice' the message of untrustworthiness, they respond by lowering their lying cost. We then prove the principal will then be forced to sharpen the incentives, sending more messages of untrustworthiness, leading to even sharper incentives, and so on.

We consider the consequences as the lying cost falls from the heights of a completely trustworthy agent. Lying costs become disruptive when they are small enough to force the abandonment of the first-best solution; large enough falls in the lying cost eventually lead to the agent earning rent, but ultimately, if this rent becomes too high, the principal will find it profit-maximizing to shut down the activities of the firm. Finally, we note that the contract offered to an agent with a low-enough-lying cost can extract rent which is increasing in the willingness to lie. Thus, appearing dishonest can be a form of rent-seeking.

The paper is structured as follows. We conclude Section 1 with a discussion of related literature. In Section 2, we model the situation where an agent's lying cost is known to the principal and show how it outperforms the classic solution for any non-zero lying cost. In Section 3, we solve the optimal menu for the case where only the distribution of the lying cost is known by the principal. In Section 4, we consider a "feedback effect," where the offer of an incentive contract lowers lying costs by communicating untrustworthiness. Section 5 concludes and flags future research by way of two conjectures. Appendix A offers a discussion on lying costs and trustworthiness. All proofs are relegated to Appendix B.

1.1. Related Literature. There are many important applications of principal-agent problems with moral hazard, and they have been extensively studied in Economics and other Management literature. To name a few initial seminal works on moral hazard models: Arrow (1970) and Spence and Zeckhauser (1978) on the theory of insurance under moral hazard; Shapiro and Stiglitz (1984) on efficient wage theories; and Alchian and Demsetz (1972), Jensen and Meckling (1976), Grossman and Hart (1982) on the theory of the managerial firm. More specifically, Jensen and Meckling (1976) argue that a conflict exists between equity owners and managers because the managers only get a fraction of the firm's profit while bearing the full cost of their own effort in enhancing the firm's profitability. We refer the reader to Laffont and Martimort (2009) and Bolton and Dewatripont (2004) for a detailed discussion of principal-agent problems with moral hazard.

The literature about hidden action (or hidden information) with lying costs is surprisingly small. In a seminal paper, Kartik (2009) studies a sender-receiver (cheap talk) model where the sender bears the cost of lying.[1] The main result of this paper is that the sender typically claims to be of a higher type than he would with complete information, and an incomplete separation among different types occurs. Mainly motivated by sharecropping models, Crocker and Morgan (1998) solve for the optimal contracts when insured

[1] In more recent work, the related concepts of "costly calibration," and "falsification" are studied by Guo and Shmaya (2021) and Perez-Richet and Skreta (2022), respectively.

individuals possess private information about their losses and can misrepresent their loss magnitudes while incurring a cost from such falsification. In a monopoly screening model, Severinov and Deneckere (2006) consider a monopolist facing consumers who have privately known demands, where a fraction of consumers are assumed to be honest (hence, when they are asked to announce their preferences, they will reveal their true preferences independent of the mechanism.)

There is a small amount of literature on mechanism design with lying costs, especially with hidden information. For instance, Deneckere and Severinov (2008) studies implementation in environments where agents have limited ability to imitate others and develops an "extended revelation principle." Kartik et al. (2014) considers full implementation (in a complete-information environment) when agents have an arbitrarily small preference for honesty, establishing a certain condition for social choice functions to be implementable. Ben-Porath and Lipman (2012) considers the implementation problem where the agents support their statements with "hard evidence," and identifies a necessary condition for the implementation. In mainly an experimental work, Charness and Dufwenberg (2006) studies the impact of communication on trust and cooperation, providing evidence that the subjects strive to live up to others' expectations so as to avoid guilt, and promises may enhance trustworthy behavior.

The growing literature on mechanism design with costly monitoring is marginally related to our paper. In this literature, there are no fundamental and explicit lying costs. Still, the principal can check an agent's information at a cost and punish the agent if a lie is detected (hence lying brings a cost within the mechanism.) The papers in this literature include Ben-Porath et al. (2014), Mylovanov and Zapechelnyuk (2017), and Halac and Yared (2020).

Our paper is a contribution to behavioral contract theory, (for a review, see Kőszegi, 2014). In addition, our paper speaks to the literature on the signaling effect of incentives by providing a theory of optimal incentives under the assumption that incentives signal untrustworthiness. For this literature, please see, for example, Sliwka (2007), Galbiati et al. (2009), and Van der Weele (2012).

To the best of our knowledge, no prior work has modeled and analyzed the simple principal-agent model in the presence of lying costs.

2. Known Lying Cost

In our generalized principal-agent model, the principal hires an agent and relies on their unobserved effort to generate stochastic revenue R or 0. The agent's outside option is zero. The agent can declare high effort (h) or low effort (l) and then do either high effort (H) or low effort (L). If the agent puts in a high effort, the probability that revenue equals R is $1-q$ (where $0<q<0.5$), and if they put in a low effort, the probability of obtaining R is q. Payment is assumed to be directly related to *declared effort* and not *actual actions.* Actions and declarations are represented $\{H \text{ or } L, h \text{ or } l\}$. The agent incurs an intrinsic lying cost $x>0$ if what is said and done diverges, and the principal knows x.[2] Effort costs $y>0$. As will become apparent, when an agent declares a high effort, the actual effort level will hinge on the relative cost of high effort versus the cost of lying.

Timing

(1) Nature chooses an x for the agent, x is revealed to both the agent and the principal.

(2) The principal chooses a, b, c, and d in the payment schedule.[3]

(3) The agent chooses actual effort and a declaration of effort. Revenue is revealed.

(4) The principal pays the agent the wage amounts.

The payments in the schedule are as follows. Model pronumerals for all sections are listed for completeness.

[2] If lying is costless (x=0), we have a special case. The agent will always claim to put in a high effort, so the declaration is meaningless. In this case, the principal's payment menu effectively becomes a payment for outcomes, and declarations are disregarded.

[3] $a, b, c, d \in \mathbb{R}_+$

Pronumeral Glossary

a^2	Payment for high revenue outcome (R) if high effort declared.
b^2	Payment for low revenue outcome (0) if high effort declared.
c^2	Payment for high revenue outcome (R) if low effort declared.
d^2	Payment for low revenue outcome (0) if low effort declared.
h, H	High effort: declared, enacted
l, L	Low effort: declared, enacted
F	cdf for lying cost (later section)
q	Probability of high revenue following low effort.
R	High revenue outcome
t	Lying cost with feedback (later section)
x	Lying cost
y	Effort cost
θ	Strength of feedback (later section)

Utility for the agent includes a deceit cost (which can be 0 or x depending on truthful revelation or lying, respectively).[4]

$$U_P = E(\pi) = E(\text{Revenue} - \text{wage}) \tag{1}$$

$$U_A = \sqrt{\text{wage}} - \text{cost of effort} - \text{deceit cost} \tag{2}$$

The actions and outcomes for the general model are shown in the top panel of Table 1. The bottom two panels become relevant as our argument progresses.

[4] The square-root representation of agent utility is without loss of generality. If we use a general $u(0) = 0$, $u' > 0$ and $u'' < 0$, all the subsequent diagrams and results would follow. The proof is available from the authors upon request.

	General Model			
	High Rev $= R$	Low Rev $= 0$		
Actions	Prob, Pay		Expected Utility	Expected Profits
$\{H,h\}$	$1-q, a^2$	q, b^2	$(1-q)a+qb-y$	$(1-q)R-[(1-q)a^2+qb^2]$
$\{L,h\}$	q, a^2	$1-q, b^2$	$qa+(1-q)b-x$	$qR-[qa^2+(1-q)b^2]$
$\{L,l\}$	q, c^2	$1-q, d^2$	$qc+(1-q)d$	$qR-[qc^2+(1-q)d^2]$
$\{H,l\}$	$1-q, c^2$	q, d^2	$(1-q)c+qd-y-x$	$(1-q)R-[(1-q)c^2+qd^2]$

	No $\{H,l\}$, c=d=0			
	High Rev $= R$	Low Rev $= 0$		
Actions	Prob, Pay		Expected Utility	Expected Profits
$\{H,h\}$	$1-q, a^2$	q, b^2	$(1-q)a+qb-y$	$(1-q)R-[(1-q)a^2+qb^2]$
$\{L,h\}$	q, a^2	$1-q, b^2$	$qa+(1-q)b-x$	$qR-[qa^2+(1-q)b^2]$
$\{L,l\}$	$q, 0$	$1-q, 0$	0	qR

	Classic Solution: $x=0$			
	High Rev $= R$	Low Rev $= 0$		
Actions	Prob, Pay		Expected Utility	Expected Profits
$\{H\}$	$1-q, a^2$	q, b^2	$(1-q)a+qb-y$	$(1-q)R-[(1-q)a^2+qb^2]$
$\{L\}$	q, a^2	$1-q, b^2$	$qa+(1-q)b$	$qR-[qa^2+(1-q)b^2]$

Table 1. Actions and Payments for Known Lying Cost

We assume $a \geq b \geq d$ and $a \geq c \geq d$. The rankings $a \geq c$ and $b \geq d$ are a reasonable extension of the 'equal pay, equal work' principle. If that is granted, then *a fortiori* less work should not be paid a greater amount. The rankings $a \geq b$ and $c \geq d$ are based on the principle that the designer does not want to reward bad outcomes with higher pay.

We now establish a simple lemma that will greatly simplify our analysis.

Lemma 1. $\{H, l\}$ will never be chosen by the agent, and in the optimal mechanism, we have $c = d = 0$. Moreover, $\{L, h\}$ is never desired by the principal.

In the light of the previous lemma, we can argue the following: the principal will set a and b to choose between $\{H, h\}$ and $\{L, l\}$ and the former will be more profitable under the following configuration of a and b:

$$(1-q)R - \left[(1-q)a^2 + qb^2\right] \geq qR \iff a \leq \sqrt{\frac{(1-2q)R}{1-q} - \frac{q}{1-q}b^2} \tag{3}$$

The principal will want high effort if R is big enough (or if q is small enough). To anticipate a remark in Section 5, the value of a chosen in the optimal menu is decreasing in the cost of lying. For a sufficiently low cost of lying, a may become high enough to violate this inequality, whereupon the principal will want to elicit low effort.

To solve this model, assuming high effort is optimal for the principal, we reflect on the logic of setting $c = d = 0$. If the principal wants high effort, they first unincentivize the declaration of low effort. Then, with the agent declaring high effort, the principal adjusts a and b to make the agent undertake her high effort to match her declaration. Thus, the principal motivates high effort subject to a participation constraint and a requirement that $\{H, h\}$ is no worse than $\{L, h\}$.

$$(1-q)a + qb - y \geq 0 \iff a \geq \frac{y - qb}{1-q} \tag{4}$$

$$(1-q)a + qb - y \geq qa + (1-q)b - x \iff a \geq b + \frac{y-x}{1-2q} \tag{5}$$

We can name the first constraint the individual rationality constraint and the second constraint the incentive constraint (for putting high effort rather than low effort, reporting h in both cases). In Figure 1 below, revenue is fixed at $(1-q)R$ and maximizing profits is the same as minimizing costs. The iso-profit (IP) lines $\pi = (1-q)R - [(1-q)a^2 + qb^2]$ are concave when a is expressed as a function of b in $a \times b$ space, with profits rising as the iso-profit curves shift towards the origin. In the same diagram the incentive constraint intercept slides down the a axis as x rises. When the intercept falls below the point denoted by R' in Figure 1, the constraint ceases to be the feasible set. In this case, the least-cost overlap of the last two inequalities becomes the intersection of them, sliding down the participation constraint until the first-best solution $a = b = y$.[5] As x grows to be higher than y, the last equation no longer binds, and the feasible set becomes the area north of the participation constraint and North-West of the line $a = b$.

[5] $\max_{a,b}(1-q)R - [(1-q)a^2 + qb^2]$s.t.$(1-q)a + qb - y = 0$
$L = (1-q)R - [(1-q)a^2 + qb^2] + \lambda[(1-q)a + qb - y] \iff a = b = y$. In the diagram, the iso-profit locus is at a tangency to the participation constraint at this point.

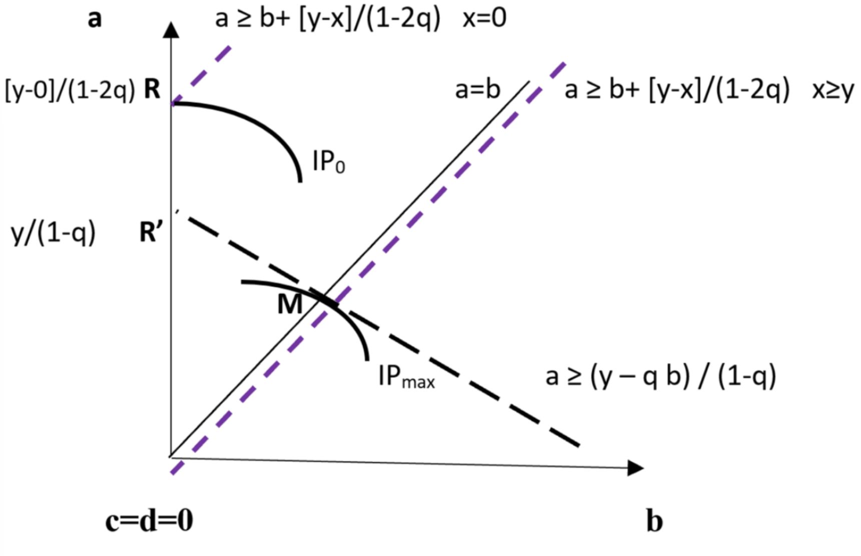

FIGURE 1. **Solution with Known Lying Cost:** Point R is the minimum-cost menu that satisfies the constraint that truthful high effort is preferred by the agent to falsely-claimed high effort when x=0. By Proposition 1 below, it is the classic principal-agent solution. Upper dashed line represents $a \geq b + \frac{y-x}{1-2q}$ for ($x = 0$.) The participation constraint is $a \geq \frac{y-qb}{1-q}$. As x rises, first the agent loses her rent (between R and R',) and for a much higher lying cost, if $x \geq y$, the first-best solution M becomes feasible. Lower dashed line represents $a \geq b + \frac{y-x}{1-2q}$ when lying is more costly than putting high effort costless ($x > y$.)

We summarize interesting takeaways from our discussions with the following Proposition.

Proposition 1. In the optimal menu, we have the following:

(1) If lying is costless, the solution (Point R) is the classic principal-agent solution.
(2) As lying costs rise above zero, hidden action rent (the vertical distance above R') is reduced.
(3) If x rises above $\frac{qy}{1-q}$, the hidden action rent becomes zero.
(4) The optimal menu motivates the agent to both put in the high effort, and, tell the truth about the effort level.

Lastly, we restate the condition for when the first-best is attainable.

Proposition 2. When lying becomes more expensive than high effort, the first-best is attainable.

In the next section, we extend our analysis to the case where the lying cost is the agent's private information.

3. Private Information Case

For the private information case, we suppose the x is non-negative and stochastic, the cumulative distribution function $F(x)$ is known to the principal, and x is known to the agent. The setup and expected utilities are identical, except the expected profits are not the same as in Section 2:

	High Rev $= R$	Low Rev $= 0$	
Actions	Prob, Pay		Expected Utility
$\{H, h\}$	$1-q, a^2$	q, b^2	$(1-q)a + qb - y$
$\{L, h\}$	q, a^2	$1-q, b^2$	$qa + (1-q)b - x$
$\{L, l\}$	q, c^2	$1-q, d^2$	$qc + (1-q)d$
$\{H, l\}$	$1-q, c^2$	q, d^2	$(1-q)c + qd - y - x$

Table 2. Utility specification: x is private information

Given x, among actions $\{H, h\}$, $\{L, h\}$, $\{L, l\}$, $\{H, l\}$, the agent will choose the action that maximizes her expected payoff. Let us denote the probability of a particular action $\{\cdots\}$ be chosen by the agent by $P(\{\cdots\})$. The principal's expected profit, ***EP***, is then given by

$$\begin{aligned}
\boldsymbol{EP} \equiv\ & P(\{H,h\})\big((1-q)R - [(1-q)a^2 + qb^2]\big) \\
&+ P(\{L,h\})\big(qR - [qa^2 + (1-q)b^2]\big) \\
&+ P(\{L,l\})\big(qR - [qc^2 + (1-q)d^2]\big) \\
&+ P(\{H,l\})\big((1-q)R - [(1-q)c^2 + qd^2]\big)
\end{aligned}$$

Given q, y and lying cost distribution F, the principal will choose a, b, c, and d to maximize ***EP***. We first establish a useful lemma where the proof is skipped since the proof has the same arguments in Lemma 1.

Lemma 2. In the optimal menu, $\{H, l\}$ will not be chosen and we have $c = d = 0$.

Given this lemma, the principal's expected profit can now be simplified by excluding $\{H, l\}$ and setting $c = d = 0$.

$$\begin{aligned}
&\boldsymbol{EP} \equiv P(\{H,h\})\left((1-q)R - [(1-q)a^2 + qb^2]\right)\\
&+P(\{L,h\})\left(qR - [qa^2 + (1-q)b^2]\right)\\
&+P(\{L,l\})qR
\end{aligned}$$

We now establish three lemmas that will be useful for our main result of this section.

Lemma 3. Among menus with $a(1-q) + bq < y$, the optimal menu is $a = b = 0$.

Lemma 4. Among menus with $a(1-q) + bq > y$, in the optimal menu, we have $b = 0$.

Lemma 5. Among menus with $a(1-q) + bq = y$, we may have optimal menus where have $b > 0$.

Now, we are ready for our main result for this section.

Proposition 3. The optimal menu is one of the following 3 options:

(1) $a = b = c = d = 0$

(2) $c = d = 0$, $b = (y - (1-q)a)/q$, and a maximizes

$$(1 - F(z))\left(R(1-q) - a^2(1-q) - b^2q\right) + F(z)\left(Rq - a^2q - b^2(1-q)\right)$$

where $z = \frac{(1-q)y-(1-2q)a}{q}$ subject to $a \geq y$

(3) $b = c = d = 0$ and a maximizes

$$\left((1 - F(z))\,(1-q) + F(z)q\right)(R - a^2)$$

for $z = y - a(1-2q)$ subject to $a > \frac{y}{1-q}$.

There are two significant differences between the known x case and the private information x case. The first one is that, in the known x case, for the optimal menu, the agent chooses high effort and declares high effort, hence the agent is always truthful. Whereas in the private information x case, for the optimal menu given in (2) and (3) in the above proposition, the agent always declares high effort but does not always choose high effort. More specifically, only agents whose lying costs are higher than z in (2) and (3) actually put in the high effort and agents with lying costs lower than z put in the low effort (and hence they are not truthful). The second important difference between the two models follows from the following proposition.

In it, we show that in the optimal menu, we have $a > b$ whenever they are not equal to 0.

Proposition 4. In the optimal menu, we can never have $a = b > 0$.

This proposition is in contrast with the result of Section 2, where we can have $a = b > 0$ in the optimal solution; more specifically $a = b = y$. Note that, for this to be the case for the known x case, we need to have $x \geq y$. In the incomplete information case, however, we assume that x is continuously distributed between 0 and ∞. Instead, if we assume the lower bound for the distribution to be y or higher, it is not difficult to see that the optimal solution would satisfy $a = b = y$ and the first-best solution would be achieved. Another related remark is that, for the private information lying cost case, as long as there is a possibility that the lying cost is lower than the high effort course (i.e. lower bound for type distribution of lying costs is smaller than y) the first-best solution cannot be achieved.

In the next section, we consider a specific feedback effect where the difference between wages communicates untrustworthiness and reduces lying costs.

4. Negative Feedback

We now use our model to consider the case of negative feedback arising from offering incentive contracts. The classic principal-agent model has been widely applied to advice-seeking from lawyers, doctors, or financiers (see Laffont and Martimort (2009) and Bolton and Dewatripont (2004)). For applications involving the reliance on experts, professionalism has been advanced as an alternative to incentive contracts. Crucially, professionalism uses a 'reasonable person' representative agent from tort law, for whom 'loyalty' makes sense in any relationship of trust, such as seeing a doctor or investing money (Lydenberg, 2014). This representative agent differs from 'homo economicus', who has a zero lying cost. Any group of experts which shift from a professional understanding of their role to a profit- or shareholder-value-maximizing one is in a theoretical sense swapping a more trustworthy representative agent for a less trustworthy one. Our generalized model can illustrate what occurs when such a cultural transformation lowers the cost of lying.[6]

We allow endogenous feedback whereby being offered an incentive contract itself communicates that the principal has opted for a 'homo economicus' view of an agency relationship rather than a 'reasonable person' view. This, in turn, is assumed to lower the cost of lying.

We consider what happens if the lying cost is no longer x, but it is given by $x+\theta(b-a)$ where $\theta \geq 0$ measure the magnitude of feedback effect, and the $(b-a)$ term implies incentive contracts (i.e. $a \geq b$) make lying cheaper, by communicating the agent is not trusted.[7]

We first consider the case when x is common knowledge, and then discuss why the incomplete information case has a similar intuition.

Firstly, we make a technical adjustment to ensure the lying cost cannot be negative and refine the definition of the lying cost as follows:

$$t \equiv max(0, x + \theta(b-a))$$

This gives us an identical payoff table to the top panel of Table 1, but with t replacing x as shown in Table 3. As before, the action $\{H, l\}$ will not be chosen (as in Lemma 1). We compare the payoffs for $\{H, h\}$ with the payoffs for $\{H, l\}$. Their difference $(1-q)a+$

[6] We provide a more detailed discussion of lying costs and trustworthiness in Appendix A.

[7] We ignore the effect of $(c-d)$ on lying cost, since in the optimal menu c and d turn out to be 0.

$qb - y) - (1-q)c - qd + y + t$ is non-negative as in the earlier proof. Crucially, the zero floor on t from our technical adjustment guarantees that the last term in the difference is non-negative.

Furthermore $c = d = 0$ in the optimal mechanism, since Lemma 1 can be applied directly. The extra effect arising from $\theta(b-a)$ does not alter the steps of the proof. We then write down the lower panel of Table 3.

	General Model			
	High Rev $= R$	Low Rev $= 0$		
Actions	Prob, Pay		Expected Utility	Expected Profits
$\{H,h\}$	$1-q, a^2$	q, b^2	$(1-q)a + qb - y$	$(1-q)R - [(1-q)a^2 + qb^2]$
$\{L,h\}$	q, a^2	$1-q, b^2$	$qa + (1-q)b - t$	$qR - [qa^2 + (1-q)b^2]$
$\{L,l\}$	q, c^2	$1-q, d^2$	$qc + (1-q)d$	$qR - [qc^2 + (1-q)d^2]$
$\{H,l\}$	$1-q, c^2$	q, d^2	$(1-q)c + qd - y - t$	$(1-q)R - [(1-q)c^2 + qd^2]$

	No $\{H,l\}$ c=d=0			
	High Rev $= R$	Low Rev $= 0$		
Actions	Prob, Pay		Expected Utility	Expected Profits
$\{H,h\}$	$1-q, a^2$	q, b^2	$(1-q)a + qb - y$	$(1-q)R - [(1-q)a^2 + qb^2]$
$\{L,h\}$	q, a^2	$1-q, b^2$	$(q+\theta)a + (1-q-\theta)b - \text{x}$	$qR - [qa^2 + (1-q)b^2]$
$\{L,l\}$	$q, 0$	$1-q, 0$	0	qR

Table 3. Actions and Payments when Feedback is present

As before, the solution involves meeting the participation constraint–that the expected utility from $\{H,h\}$ is nonnegative–and ensuring that $\{H,h\}$ is no less attractive than $\{L,h\}$. The equivalents of (4) and (5) are:

$$(1-q)a + qb - y \geq 0 \iff a \geq \frac{y - qb}{1-q} \tag{6}$$

$$(1-q)a + qb - y \geq (q+\theta)a + (1-q-\theta)b - \text{t} \iff a \geq b + \frac{y - x}{1 - 2q - \theta} \tag{7}$$

where (7) is derived assuming t is positive. The figure for the solution is simply Figure 1 with a smaller denominator in the inequality on the RHS in 7.

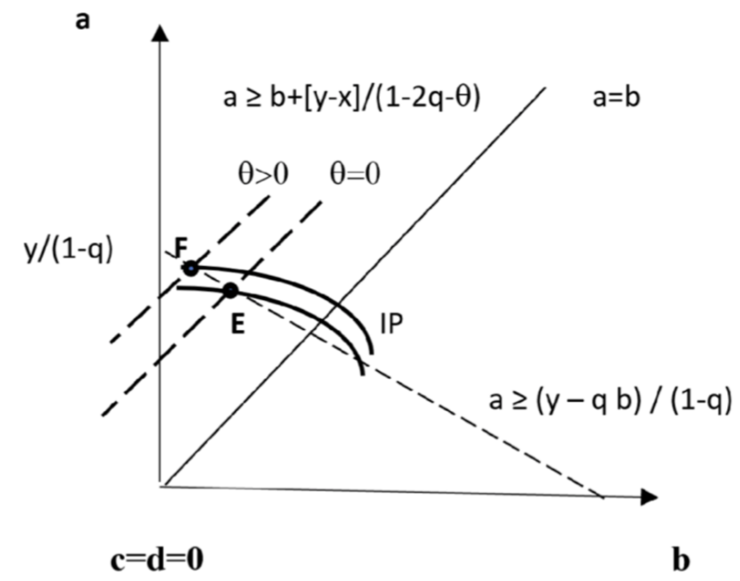

FIGURE 2. **Solution with Known Lying Cost and Feedback:** In an initial equilibrium E, x is small enough to bind on the solution so that first-best ($a = b = y$) is infeasible. If θ then rises above 0, (so there is feedback,) a must increase, and b must decrease to prevent lying. The final solution F has even sharper incentives because of this feedback.

A number of interesting points follow immediately from Figure 2 when we consider θ rising from 0. A rising θ will lift the dashed line, shrinking the feasible set. This means that in any solution without feedback ($\theta = 0$), the introduction of feedback will lead to greater incentivization, not less, as the dashed line heads northwest by parallel shifts.

This is a surprising and interesting result. One might have thought the designer would seek to make a closer to b to avoid the effects of any feedback, but unfortunately, this is not possible. For any solution, apart from $a = b = y$, a rise in θ makes any incentive contract ($a > b$) collapse as $\{L, h\}$ becomes optimal—the agent still says h but switches from doing H to L. The designer has no choice but to raise the rewards for high effort, so the agent who says h actually does H.

When t falls to zero (i.e. $x + \theta(b - a)$ is negative) this is equivalent to x being zero in the previous model without feedback, which is, in turn, equivalent to the classic principal-agent solution of Figure 1 (Point R).

When the participation constraint binds, the solution for a and b, written in terms of deviation from the first-best $a = b = y$, makes it clear that a rising θ (the presence of feedback) will sharpen the incentives away from first-best.

$$a = y + \frac{q}{1 - 2q - \theta}(y - x) \tag{8}$$

$$b = y - \frac{1 - q}{1 - 2q - \theta}(y - x) \tag{9}$$

Naturally, this nests the previous solution when $\theta = 0$.

We can discern a 'vicious circle,' or a negative feedback loop. If we start with a given solution with $a > b$, and then θ rises from zero, there will be a need to raise a relative to b along the participation constraint. But this, in turn, will make $\{L, h\}$ optimal in the absence of further adjustment. To address this, the designer moves the incentives further North West along the participation constraint, each round requiring a little more subsequent adjustment to discourage lying, until settling at the equations 8 and 9.

We summarize the above discussions in the following Proposition (where the proof is skipped since the proof follows from the above discussions.)

Proposition 5. In the optimal menu, the difference between (root) wage differences for high revenue and low revenue case (i.e. $a - b$) is higher for a positive θ as compared to no feedback case ($\theta = 0$.) Hence, feedback results in more discriminative offerings.

This vicious circle cautions against the indiscriminate offering of incentive contracts if, by communicating to the recipients that they are untrustworthy, it comes to pass that there is a greater propensity to lie. The ethical drift within the international banking industry prior to the GFC suggests this is a realistic possibility (Salz and Collins, 2013).

We can argue that the same results would hold for the incomplete information case. Specifically, we can follow the same steps of Section 5 simply by using t instead of x. For the feedback case, in Proposition 3, the cases (2) and (3) will be adjusted as follows:

(2) $c = d = 0$, $b = (y - (1-q)a)/q$, and a maximizes

$$(1 - F(z))\left(R(1-q) - a^2(1-q) - b^2 q\right) + F(z)\left(Rq - a^2 q - b^2(1-q)\right)$$

where $z = \frac{(1-q)y-(1-2q-\theta)a}{q}$ subject to $a \geq y$

(3) $b = c = d = 0$ and a maximizes

$$\left((1 - F(z))(1-q) + F(z)q\right)(R - a^2)$$

for $z = y - a(1 - 2q - \theta)$ subject to $a > \frac{y}{1-q}$.

Now, when we compare $\theta > 0$ to the no-feedback case, it is not difficult to see that a will be higher for both (2) and (3), whereas b will remain the same in (2) and be smaller in (3). Hence, we can establish that feedback results in more discriminative offerings, even for the private information case.

Having laid out the generalized model for both public and private x, and the potential for feedback, the next section lays out some other ideas the model could investigate.

5. Discussion and Conclusion

We offer two considerations in this final section to give a sense of future research. First, cultures of untruthfulness may contribute to economic decline, and second, appearing dishonest is rent-enhancing, and, therefore could, if pursued strategically, be a form of rent-seeking. In this section, x reverts back to the true lying cost.

Our first consideration is about economic decline, and it relates directly to the different solutions offered by the general model, as x falls. Tracing through this decline in Figure 1 in terms of the incentive constraint, we begin South East of the first-best optimum, where the fully optimal solution is feasible. In this region, the principal doesn't have to believe the agent is Kantian (unable to lie, with an infinitely large x). Still, they do believe the agent will not lie under 'standard' circumstances. Or, as we noted in Section 3, even if x is unknown to the principal, agents are far from their breaking point. Either way, the level of trust is high enough that the principal is prepared to pay a fixed amount and rely on the reported efforts by the agent.

As x falls to the point of disrupting the first-best, the principal can no longer rely on the trustworthiness of the agent. As in the solutions offered above, the principal pays on claims of the work having been done well, but makes sure bad outcomes are penalized to some degree, so that these claims of high effort are likely to correspond to reality. If writing these contracts is not feasible, or if the principal faces the stochastic and unknown x of the previous section, sometimes the agent will falsely claim to have put in high effort.

Eventually, if x declines further, the principal will have to pay a sufficiently high payment to the agent that it becomes profit-maximizing for the principal to no longer incentivize high effort. The principal offers $a = b = 0$, and the economy produces output randomly without high effort. The specter of falling x can be motivated interestingly by a change in, say, a political system. One account of East German ethical norms was that they encouraged relative untruthfulness, presumably because of the political and economic environment (Ariely et al., 2015).

Our second consideration highlights the dangers of making theory too narrow when applied in a policy environment: choosing a special case among competing models can advantage certain economic actors. An interesting feature of the classic Jensen and Mecking contract ($x = 0$) is that it allows the agent to earn rent from appearing untrustworthy. That is, agents in the 1980s, on the cusp of deregulation, received a commercial advantage when the standard principal-agent solution was accepted as ‘the’ way to frame contracts rather than adopting the alternative solution of professionalism. In our model, professionalism would be modeled by raising x through the development of professional culture. This would open the way for the first-best solution, or at least enable a solution that does not deliver rent to the agent. We leave it to economic historians to investigate if this was a real motivation for the promotion of bonus schemes for super-managers in the closing decades of the 20th century.

It is well-known that those in charge of setting compensation rates found justification for lucrative schemes (Piketty, 2014). The discussion here can go further by asking if the framework of bonus schemes paid managers rent, aided and abetted by a theory that discarded the 'reasonable person' representative agent of professionalism in favor of 'homo economicus'.

Our general principal-agent model suggests that an agent has an interest in convincing a principal that they are untrustworthy at the time the terms of a contract are decided. Moreover, the incentive for rent-seeking is non-linear. Over some range of moderately low x's, the optimal contract will cost the principal profits, but will not take the agent off their participation constraint. At some lower value of x, however, the optimal contract delivers rent to the agent. Seen in this light, the implicit choice of $x = 0$ for 'homo economicus' was, if not rent-seeking, certainly rent-creating.

To conclude, this paper has made a simple point with a simple model: hidden information is only valuable if someone is prepared to hide it by lying. Yet reliable (i.e. unincentivized) truthtelling is a feature of the real world, implying that sometimes the value of hidden information remains unrealized. We, therefore, propose that truth-telling proclivity be made explicit in all models involving communication, because the revelation of hidden information, when it does occur, attenuates both rent and inefficiency.

Furthermore, nothing in our model is utopian about truthfulness. We have parameterized the degree of truthtelling by x, so there is no presumption that the truth is often, or seldom, spoken in any particular context. Our generalized principle-agent model can therefore be applied in every context where the standard zero-lying-cost case has paved the way.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond, "Preferences for truth-telling," *Econometrica*, 2019, *87* (4), 1115–1153.

Alchian, Armen A and Harold Demsetz, "Production, information costs, and economic organization," *The American economic review*, 1972, *62* (5), 777–795.

Ariely, Dan, Ximena Garcia-Rada, Lars Hornuf, and Heather Mann, "The (true) legacy of two really existing economic systems," 2015.

Arrow, Kenneth J, "Essays in the theory of risk-bearing," Technical Report 1970.

Bacharach, Michael, Gerardo Guerra, and Daniel John Zizzo, "The self-fulfilling property of trust: An experimental study," *Theory and Decision*, 2007, *63*, 349–388.

Bauman, Yoram and Elaina Rose, "Selection or indoctrination: Why do economics students donate less than the rest?," *Journal of Economic Behavior & Organization*, 2011, *79* (3), 318–327.

Ben-Porath, Elchanan and Barton L Lipman, "Implementation with partial provability," *Journal of Economic Theory*, 2012, *147* (5), 1689–1724.

______, Eddie Dekel, and Barton L Lipman, "Optimal allocation with costly verification," *American Economic Review*, 2014, *104* (12), 3779–3813.

Bolton, Patrick and Mathias Dewatripont, *Contract theory*, MIT press, 2004.

Büchner, Susanne, Luis G González, Werner Güth, and M Vittoria Levati, "Incentive contracts versus trust in three-person ultimatum games: an experimental study," *European Journal of Political Economy*, 2004, *20* (3), 673–694.

Charness, Gary and Martin Dufwenberg, "Promises and partnership," *Econometrica*, 2006, *74* (6), 1579–1601.

Crocker, Keith J and John Morgan, "Is honesty the best policy? Curtailing insurance fraud through optimal incentive contracts," *Journal of Political Economy*, 1998, *106* (2), 355–375.

Deneckere, Raymond and Sergei Severinov, "Mechanism design with partial state verifiability," *Games and Economic Behavior*, 2008, *64* (2), 487–513.

der Weele, Joel Van, "The signaling power of sanctions in social dilemmas," *The Journal of Law, Economics, & Organization*, 2012, *28* (1), 103–126.

Erat, Sanjiv and Uri Gneezy, "White lies," *Management Science*, 2012, *58* (4), 723–733.

Fligstein, Neil and Alexander F Roehrkasse, "The causes of fraud in the financial crisis of 2007 to 2009: Evidence from the mortgage-backed securities industry," *American Sociological Review*, 2016, *81* (4), 617–643.

Galbiati, Roberto, Karl H Schlag, and Joël J van der Weele, "Can sanctions induce pessimism? An experiment," *An Experiment* (*March 24, 2009*), 2009.

Gneezy, Uri, Agne Kajackaite, and Joel Sobel, "Lying aversion and the size of the lie," *American Economic Review*, 2018, *108* (2), 419–53.

______ **and Aldo Rustichini**, "A fine is a price," *The journal of legal studies*, 2000, *29* (1), 1–17.

Grossman, Sanford J and Oliver D Hart, "Corporate financial structure and managerial incentives," in "The economics of information and uncertainty," University of Chicago Press, 1982, pp. 107–140.

Guerra, Gerardo and Daniel John Zizzo, "Trust responsiveness and beliefs," *Journal of Economic Behavior & Organization*, 2004, *55* (1), 25–30.

Guo, Yingni and Eran Shmaya, "Costly miscalibration," *Theoretical Economics*, 2021, *16* (2), 477–506.

Halac, Marina and Pierre Yared, "Commitment versus flexibility with costly verification," *Journal of Political Economy*, 2020, *128* (12), 4523–4573.

Hill, Claire A and Richard W Painter, "Better bankers, better banks," in "Better Bankers, Better Banks," University of Chicago Press, 2015.

______ **and** ______, *Better bankers, better banks: Promoting good business through contractual commitment*, University of Chicago Press, 2020.

Hogan, W, "Prospects and Challenges for Australia's Financial System', Centre for Policy Market and Design Workshop on Banking," *Health and Education, University of Technology Sydney, November*, 2018, *22*, 2019–01.

Jensen, Michael C, "Integrity: Without it nothing works," *Rotman Magazine: The Magazine* *of the Rotman School of Management*, 2009, pp. 16–20.

——— **and William H Meckling**, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *Journal of financial economics*, 1976, *3* (4), 305–360.

Kartik, Navin, "Strategic communication with lying costs," *The Review of Economic Studies*, 2009, *76* (4), 1359–1395.

———**, Olivier Tercieux, and Richard Holden**, "Simple mechanisms and preferences for honesty," *Games and Economic Behavior*, 2014, *83*, 284–290.

Khurana, Rakesh, "From higher aims to hired hands," in "From Higher Aims to Hired Hands," Princeton University Press, 2010.

Kőszegi, Botond, "Behavioral contract theory," *Journal of Economic Literature*, 2014, *52* (4), 1075–1118.

Laffont, Jean-Jacques and David Martimort, "The theory of incentives," in "The Theory of Incentives," Princeton university press, 2009.

Lydenberg, Steve, "Reason, rationality, and fiduciary duty," *Journal of business ethics*, 2014, *119* (3), 365–380.

Martin, Iain, *Crash Bang Wallop: The Inside Story of London's Big Bang and a Financial Revolution that Changed the World*, Hachette UK, 2016.

Menzies, Gordon, Donald Hay, Thomas Simpson, and David Vines, "Restoring Trust in Finance: From Principal–Agent to Principled Agent," *Economic Record*, 2019, *95* (311), 497–509.

Morris, Nicholas and David Vines, *Capital failure: Rebuilding trust in financial services*, OUP Oxford, 2014.

Mylovanov, Tymofiy and Andriy Zapechelnyuk, "Optimal allocation with ex post verification and limited penalties," *American Economic Review*, 2017, *107* (9), 2666–94.

Perez-Richet, Eduardo and Vasiliki Skreta, "Test design under falsification," *Econometrica*, 2022, *90* (3), 1109–1142.

Piketty, Thomas, "Capital in the twenty-first century," in "Capital in the twenty-first century," Harvard University Press, 2014.

Salz, Anthony and Russel Collins, "Salz review: An independent review of Barclays' business practices," *Barclays PLC*, 2013.

Samuelson, Paul A, "Altruism as a problem involving group versus individual selection in economics and biology," *The American Economic Review*, 1993, *83* (2), 143–148.

Severinov, Sergei and Raymond Deneckere, "Screening when some agents are nonstrategic: does a monopoly need to exclude?," *The RAND Journal of Economics*, 2006, *37* (4), 816–840.

Shapiro, Carl and Joseph E Stiglitz, "Equilibrium unemployment as a worker discipline device," *The American Economic Review*, 1984, *74* (3), 433–444.

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

Spence, Michael and Richard Zeckhauser, "Insurance, information, and individual action," in "Uncertainty in Economics," Elsevier, 1978, pp. 333–343.

Appendix A: A Discussion on Lying Costs and Trustworthiness

Our model innovation of explicit lying costs allows us to characterize the collapse in trustworthiness that heralded *inter alia* the 2008 Global Financial Crisis. Echoing the Jensen quote in our opening section, Morris and Vines (2014) describes how the pre-1980s sense of duty in the financial services industry was dismantled across many jurisdictions. In its place, a shallow form of trustworthiness developed, based on the cost-benefit analysis of repeated interactions. As Jensen implied, however, true trustworthiness means acting in the interests of others when a cost-benefit analysis *fails*. Morris and Vines acknowledge the sometimes patchy extent to which the 'gentlemen bankers' (in the UK) lived out their fiduciary duty, but they argue convincingly that this duty was not to be gainsayed. After three decades of deregulation, the landscape, described in Hill and Painter (2015), had changed for the worse. One key development was the evolution of compensation schemes used in the financial services industry.

From the 1980s onwards, the principal-agent model had been increasingly applied to advice-seeking across a wide range of contexts—from lawyers, doctors, or financiers (see Laffont and Martimort (2009) and Bolton and Dewatripont (2004)). Prior to that, situations involving the reliance on experts had drawn more on notions of professionalism. To this day, as noted by Büchner et al. (2004), professionalism remains a common ethos in the public sector. Notions of professionalism hinge on a 'reasonable person' representative agent from tort law, for whom 'loyalty' makes sense in any relationship of trust, such as seeing a doctor or investing money (Lydenberg, 2014). This representative agent differs from 'homo economicus', who, as demonstrated in the body of the paper, has a zero lying cost. Any group of experts which shift from a professional understanding of their role to a profit- or shareholder-value-maximizing one is, in a theoretical sense, swapping a more trustworthy representative agent for a less trustworthy one.

Our generalized principal-agent model allows a description of this change in an analytically powerful way. Suppose that for the case of a known lying cost, the principal makes a mistake about the moral character of the agent. Keeping the same notation, the true lying cost, X, which would allow a first-best contract, is greater than the principal's belief of the lying cost, x, which disrupts the first-best contract.

This can be motivated by the observation that economists, at least in caricature, embrace a somewhat skeptical model of people's motivations.

> *Mesmerised by Homo Economicus, who acts solely on egoism, economists shy away from altruism, almost comically. Caught in a shameful act of heroism, they aver "Shucks, it was only enlightened self-interest"*. Samuelson (1993) pg. 143.

Furthermore, literature represented by, say, Bauman and Rose (2011) shows that students of economics are less pro-social and that at least some of this is a training effect rather than a selection effect.

If a hypothetical principal receives such training, either directly or via the 1980s workplace 'liberalization' culture, they might interpret good evidence that the agent's lying cost is X incorrectly and claim that the lying cost is γx where γ is strictly less than one. Unfortunately, the agent who receives a contract menu based on an overly pessimistic character assessment thus receives a 'message' about the principal's view of them.

It is far from implausible that such a message could have a morally deleterious effect. That is, either a sense of offense from being regarded as untrustworthy may resolve the agent to live up (or, more precisely, 'down') to the principal's expectation, or, the offended agent may leave the industry, presumably paving the way for a less trustworthy replacement adversely-selected from the general pool of financial professionals. The model language for either change would be that a represenative agent would have a lower true lying cost X, leading to a lower perceived lying cost x.

In the main text, we build this conjecture into the model by assuming that the offering of an incentive contract lowers the cost of lying. For modeling simplicity, we keep x as the principal's beliefs about the agent's lying cost, which is a fixed fraction of the true lying cost. There are other possible narratives of feedback, driven by a general signal that the principal has embraced a form of 'homo economicus' rather than a 'reasonable person', or, we could have assumed that the offering of an incentive contract somehow nudges the agent from a moral frame towards a completely commercial frame as they exercise their employment duties (Gneezy and Rustichini, 2000).

Thus the functional form in the main text covered a range of scenarios in the simplest possible way, by asserting that the lying cost is no longer x but is given by $x+\theta(b-a)$ where $x, \theta \geq 0$ and θ measures the magnitude of feedback effect. The $(b-a)$ term implies incentive contracts (i.e. $a \geq b$) make lying cheaper, by communicating the agent is not trusted.

Appendix B: Omitted Proofs

Proof of Lemma 1. First, since $a \geq c$ and $b \geq d$, $\{H, h\}$ weakly dominates $\{H, l\}$ and so the latter is never chosen. Next, we can argue that $c = d = 0$ using the method of contradiction. Suppose contrariwise that in the optimal mechanism, c and d are both positive. However, the payment menu $((a-d)^2, (b-d)^2, (c-d)^2, 0)$ leaves the relative payment rankings of $\{H, h\}$, $\{L, h\}$ and $\{L, l\}$ unchanged relative to (a^2, b^2, c^2, d^2) but lowers costs. Therefore c and d cannot both be positive.

We next suppose that d is zero, but c is strictly positive (recall that c cannot be less than d). However, the payment menu $((a-cq)^2, (b-cq)^2, 0, 0)$ leaves the relative payment rankings unchanged relative to $(a^2, b^2, c^2, 0)$ but lowers costs. Therefore c cannot be strictly positive without entailing a contradiction.

Next, we easily establish that $\{L, h\}$ is never desired by the principal: inspection of the second panel of Table 1 shows the expected revenue is the same as $\{L, l\}$ but the latter choice involves no wage paid to the agent, so it is preferable for the principal.

Having established that we can omit $\{H, l\}$ and set $c = d = 0$ while solving the optimal menu, we obtain the actions and payoffs in the second panel of Table 1.

□

Proof of Proposition 1. Regarding statement 1, the classic solution to the principal-agent problem is given by point R in the diagram. Consider a parallel problem to our setup where the principal only pays for revenue outcomes: a^2 for high revenue and b^2 for low revenue. If the principal wants high effort, they must ensure that the participation constraint is met (individual rationality constraint), and that high effort pays at least as well as low effort (incentive constraint), viz.;

$$(1-q)a + qb - y \geq 0 \iff a \geq \frac{y - qb}{1-q} \tag{10}$$

$$(1-q)a + qb - y \geq qa + (1-q)b \iff a \geq b + \frac{y}{1-2q} \tag{11}$$

Since the overlap of these two sets is North of the upper dashed line in Figure 1, the solution is identical to the solution of the general model with $x = 0$. The intuition is that if x=0, the agent will declare h regardless of their effort, since h attracts a higher payment and there is no penalty for lying. The principal knows this, and so ignores any declaration, meaning that the principal only pays according to revenue outcomes. This makes the last panel of Table 1 a special case of the middle panel.

We note that in the classic solution, the agent earns rent shown as $R - R'$ in the diagram. This is a consequence of the assumption that b cannot be negative. If this were not so, a solution could be obtained at the intersection of the above two inequalities (written as equalities) with b being strictly negative. Conversely, any classic solution with a floor on b which is strictly positive will earn more rent than in the figure.

Regarding statement 2, by inspection of Figure 1, as x rises the feasible set grows and higher profits can be attained by lowering a. Intuitively, as x rises, putting in high effort after declaring h becomes more attractive, allowing the principal to lower the incentives for H.

Regarding statement 3, when x rises to $\frac{qy}{1-q}$ the high effort constraint intercept drops to $\frac{y}{1-q}$, which is the participation intercept. Further falls in x make the participation constraint bind, and the solution becomes the intersection of the two constraints–the contracts offered slide South East down the participation constraint. Since all solutions for higher x are on the participation constraint, the rent is zero.

Finally, statement 4 is obvious since in the optimal menu putting low effort (i.e. choosing $\{L, l\}$ or $\{L, h\}$) gives weakly worse utility than putting high effort and declaring high effort (i.e. choosing $\{H, h\}$.)

□

Proof of Proposition 2. In Figure 1, when $y - x$ is zero (lying cost equals effort cost) the intersection of the constraints (and therefore the solution) is $a = b = y$. But this is the first-best. Once x is greater than y, the iso-profit line is at a maximum with $a = b = y$ and further expansions in the feasible set in Figure 1 are irrelevant. □

Proof of Lemma 3. When $a(1-q) + bq < y$, the agent will never choose $\{H, h\}$ as it results in a negative payoff whereas she can secure a 0 payoff by chosing $\{L, l\}$ instead. If this is the case, there is no reason for the principal to pay the agent any positive amount. □

Proof of Lemma 4. To prove this, by the method of contradiction, suppose (a, b) where $a \geq b > 0$ is optimal among menus with $a(1-q) + bq > y$. Now, consider the alternative menu $(a-\epsilon, b-\epsilon)$ where ϵ is small enough (more specifically $\epsilon > 0$, $(a-\epsilon)(1-q)+(b-\epsilon)q > y$ and $b > \epsilon$). For this alternative menu, no agent will chose $\{L, l\}$ and since agent's comparision between $\{H, h\}$ and $\{L, h\}$ is unaffected, this menu results in the same selection with (a, b) with strictly lower costs. A contradiction. □

Proof of Lemma 5. When $a(1-q)+bq = y$, the agent is indifferent between $\{H, h\}$ and $\{L, l\}$. We assume that agent choses $\{H, h\}$ in this case. Whether the agent choses $\{H, h\}$ or $\{H, l\}$ depends on the agent's lying cost. If $qa + (1-q)b - x < y$, then the agent choses $\{H, h\}$; otherwise she choses $\{H, l\}$. Let us denote $z = qa + (1-q)b = \frac{(1-q)y-(1-2q)a}{q}$. Then, the expected utility of the principal is given by

$$(1 - F(z)) \left(R(1-q) - a^2(1-q) - b^2 q\right) + F(z) \left(Rq - a^2 q - b^2(1-q)\right)$$

To confirm, numerical solution for the specification of $q = 0.2$, $y = 0.5$, $R = 2$ and $F(x) = x^3$ is $b \approx 0.17893$ and $a \approx 0.55027$. □

Proof of Proposition 3. Lemma 3 shows that the solution in (1) above maximizes the principal's payoff among menus with $a(1-q) + bq < y$; Lemma 4 shows that the solution in (2) above maximizes the principal's payoff among menus with $a(1-q)+bq = y$, and Lemma 5 shows that the solution in (3) above maximizes the principal's payoff among menus with $a(1-q)+bq > y$.

Moreover, we can numerically show that (1) is optimal when $q = 0.45$, $y = 0.9$, $R = 1$, $F(x) = x$; (2) is optimal when $q = 0.2$, $y = 0.5$, $R = 2$, $F(x) = x^3$; and (3) is optimal when $q = 0.2$, $y = 0.5$, $R = 5$ and $F(x) = x$. □

Proof of Proposition 4. Suppose we have $a = b > 0$ at the optimal menu. Then we know that optimal menu satisfies $a(1-q)+bq = y$. This is because, for a$(1-q)+bq < y$ optimal solution satisfies $a = b = 0$ and for $a(1-q)+bq < y$ optimal solution satisfies $a > b = 0$.

For $a(1-q)+bq = y$, when we take the first order derivative with respect to a and evaluate it $a = y$ (this is because when $a(1-q)+bq = y$ and $a = b$, we have $a = y$), we get $-2F(y)y\frac{1-2q}{1-q}$, which is always negative. Hence, $a = b = y$ cannot be optimal it will be strictly better to chose a slightly greater than y (and b slightly smaller than y.) □