Agency Problems

Luca Anderlini  
(Georgetown University)  
Leonardo Felli  
(London School of Economics)

July 2006

1. Introduction

Within modern economic analysis, early recognition of the importance of agency problems goes back to at least Marschak (1955), Arrow (1963) and Pauly (1968). These early works are followed by the classical contributions of Mirrlees (1975), Holmström (1979), Shavell (1979) and Grossman and Hart (1983).

The canonical form of the Principal-Agent problem still in use crystallizes in Holmström (1979) and Grossman and Hart (1983). A risk-neutral Principal $\mathcal{P}$ hires a risk-averse Agent $\mathcal{A}$. Both actors are necessary to generate output, which depends stochastically on $\mathcal{A}$’s actions. These are generally referred to as “effort” ($e$) and, crucially are not observable by $\mathcal{P}$ or any third party like a Court. In jargon, effort is neither observable nor verifiable, and hence no contractual arrangements can depend on $e$.¹ The interests of $\mathcal{P}$ and $\mathcal{A}$ are not aligned because $e$ causes disutility to $\mathcal{A}$.

$\mathcal{P}$ makes a take-it-or-leave-it offer of a contract to $\mathcal{A}$ that specifies a schedule of output-contingent wages. $\mathcal{P}$’s offer is rejected unless it meets $\mathcal{A}$’s individual rationality constraint (henceforth $IR$), stating that $\mathcal{A}$’s expected utility cannot be less than that yielded by his next best alternative employment. In addition, the problem may or may not include an explicit limited liability constraint (henceforth $LC$) stating that, regardless of output, $\mathcal{A}$’s wage cannot go below a given level. After a contract is signed, $\mathcal{A}$ chooses $e$, then the uncertain output is realized, and finally payments are made according to the contract.

In the canonical model there is a trade-off between insurance and incentives. Optimal risk-sharing would require $\mathcal{P}$ to insure $\mathcal{A}$ against output uncertainty. However, doing so would leave $\mathcal{A}$ without any incentives to exert effort: $\mathcal{A}$ would be guaranteed a constant wage and hence would choose that $e$ which gives minimal disutility. Typically, $\mathcal{P}$’s choice is instead to offer a contract that does not fully insure $\mathcal{A}$, so as to give him incentives to exert effort. The contract compensates $\mathcal{A}$ for the risk he bears in order to satisfy the $IR$ (and possibly the $LC$). If $e$ is sufficiently productive in the stochastic technology, $\mathcal{P}$’s expected profit increases as a result. The need to generate effort via incentives yields an agency problem. The equilibrium contract may be far from the “first best” world in which a social planner can choose $e$ at will. A lower than “socially efficient” $e$ is selected and $\mathcal{A}$ is not fully insured.

¹Anderlini and Felli (1998) consider a Principal-Agent problem in which $e$ is in principle contractible, but where the equilibrium contract does not include it because of complexity considerations arising from the difficulties of describing it.

---

*This essay was prepared as an entry for the 2nd Edition of the New Palgrave Dictionary of Economics. We are grateful to Larry Blume, Steven Durlauf and Maitreesh Ghatak for useful comments.
When both $\mathcal{P}$ and $\mathcal{A}$ are risk-neutral, an agency problem also arises if the LC binds (and typically the IR does not).\(^2\) In this case in order to give $\mathcal{A}$ incentives $\mathcal{P}$ can only pay him more when output indicates that effort is higher. This drives a wedge between $\mathcal{P}$’s marginal cost for increased $e$ and its social marginal cost. This in turn dictates that the equilibrium contract will differ from the first best, and a “second-best” “constrained-inefficient” outcome obtains.

Because of its tractability, the case in which both $\mathcal{P}$ and $\mathcal{A}$ are risk-neutral and the LC binds while the IR does not is a good benchmark to illustrate the mechanics of the problem and some of the more recent developments of the theory.

2. A Simple Benchmark

$\mathcal{P}$ hires $\mathcal{A}$ to carry out a task that requires unobservable non-contractible effort $e \in [0, 1]$. $\mathcal{A}$’s effort determines the probability that the task is successful in generating output. Output equals 1 with probability $e$ and 0 with probability $1 - e$. Output is observable and contractible. First, $\mathcal{P}$ offers a contract to $\mathcal{A}$, then $\mathcal{A}$ accepts or rejects it. After a contract is signed, $\mathcal{A}$ chooses $e$.

A contract is a pair of reals $(w_1, w_0)$, with the first being the wage (in units of output) that $\mathcal{P}$ pays $\mathcal{A}$ if output is 1, and the second being the wage if output is 0. Importantly, $\mathcal{A}$ has limited liability. He cannot be paid a negative wage in any state of the world. This generates the two LCs $w_1 \geq 0$ and $w_0 \geq 0$.

Both $\mathcal{P}$ and $\mathcal{A}$ are risk-neutral, and $\mathcal{A}$ dislikes effort which generates disutility $e^2/2$. Given $(w_1, w_0)$ and $e$, $\mathcal{P}$’s payoff is $e(1 - w_1) - (1 - e)w_0$, while $\mathcal{A}$’s is given by $ew_1 + (1 - e)w_0 - e^2/2$. The outside options of both $\mathcal{P}$ and $\mathcal{A}$ are normalized to zero, so that in equilibrium both expected payoffs must be non-negative. These are the IRs.

Given $(w_1, w_0)$, $\mathcal{A}$’s choice of $e$ is immediately computed as $e = w_1 - w_0$, this is the incentive constraint (henceforth IC) of the agent. If both $w_0$ and $w_1$ are lowered by the same amount $e$ does not change. Hence in equilibrium $w_0 = 0$ and $e = w_1$. Taking into account IC, $\mathcal{P}$ maximizes $e(1 - e)$. Therefore, in equilibrium, $e = w_1 = 1/2$. Hence $\mathcal{P}$’s equilibrium payoff is $\Pi^P = 1/4$, while $\mathcal{A}$’s is $\Pi^A = 1/8$, so that the IR does not bind for either $\mathcal{A}$ or $\mathcal{P}$.

If a social planner were able to choose $e$ at will, this would be chosen so as to maximize $e - e^2/2$, expected output minus cost of effort. So the first-best level of effort is $e = 1$. In this hypothetical world, $\Pi^P + \Pi^A = 1/2$, while in equilibrium $\Pi^P + \Pi^A = 3/8$. This gap is the result of the agency problem; $\mathcal{A}$ is only motivated by the difference $w_1 - w_0$. Because of limited liability, the only way for $\mathcal{P}$ to motivate $\mathcal{A}$ is to raise $w_1$. This makes $\mathcal{A}$’s effort too costly at the margin for $\mathcal{P}$: the (expected) cost of effort $e$ is $w_1e = e^2$, so that the marginal cost is $2e$. This exceeds the

\(^2\)If the reverse is true, then giving incentives to $\mathcal{A}$ has no cost since he does not mind risk and the IR binds on his expected payoff. In fact in this case, the “social optimum” coincides with the “constrained social optimum” in which a social planner can choose $e$, but only subject to giving the appropriate incentives to $\mathcal{A}$.
social marginal cost, which is $\partial/\partial e \left[ e^2/2 \right] = e$, thus inducing an inefficient second-best outcome.

3. Multi-Tasking

Starting with Holmström and Milgrom (1991), the theory evolved to encompass the multi-tasking case in which $\mathcal{A}$ has to carry out multiple tasks that affect output. Some of the insights can be conveyed adapting the simple benchmark model above.

$\mathcal{A}$ now has two tasks; one is “standard” ($S$) and one is “noisy” ($N$). He chooses two effort levels: $e_S$ and $e_N$, both in $[0, 1]$. Choosing $(e_S, e_N)$ costs $\mathcal{A}$ a disutility of $(e_S^2 + e_N^2)/4$. The two tasks are perfect complements in the stochastic technology. Given $(e_S, e_N)$, output equals 1 with probability $\min\{e_S, e_N\}$, and 0 with probability $1 - \min\{e_S, e_N\}$. As in the benchmark, $\mathcal{P}$’s payoff is expected output, minus expected wage, while $\mathcal{A}$’s payoff equals his expected wage, minus the disutility of effort. The $LC$ and $IR$ are as before.

Task $N$ is noisier than task $S$ in the following sense. Output is not contractible. Instead, each task yields a binary signal that can be contracted on. The signal $\sigma_S$ for the $S$ task is equal to 1 with probability $e_S$, and 0 with probability $1 - e_S$. The signal $\sigma_N$ for the $N$ task is equal to 1 with probability $[e_N p + (1 - e_N)(1 - p)]$ and equal to 0 with the complementary probability, with $p \in [1/2, 1]$. So, if $p = 1/2$ then $\sigma_N$ contains no information about $e_S$, while if $p = 1$, the signals $\sigma_S$ and $\sigma_N$ are equally informative about the respective tasks.

Because of the signal structure, a contract is now a quadruple of wages $(w_{S1}, w_{S0}, w_{N1}, w_{N0})$, one for each task, and for each possible value of the corresponding signal. As in the benchmark, in equilibrium we must have $w_{S0} = w_{N0} = 0$. Given $(w_{S1}, w_{S0}, w_{N1}, w_{N0})$, the ICs pin down $e_S$ and $e_N$ as satisfying $e_S = 2 w_{S1}$ and $e_N = 2 w_{N1}(2p - 1)$. Maximizing $\mathcal{P}$’s profit using these restrictions gives that in equilibrium $e_S = e_N = \max\{0, 1/2 - (1 - p)/(8p - 4)\}$. When $p = 1$ this model yields the same first best and the equilibrium payoffs as the benchmark above. When $p = 3/5$ or less then $e_S = e_N = 0$.

The literature highlights some features of the equilibrium for values of $p \in [1/2, 1]$. As $p$ decreases, so that task $N$ becomes more noisy, two changes occur. In equilibrium, $e_N$ decreases. This is not very surprising, given the increased noise. What is less straightforward is that $e_S$ decreases as well: increased noise yields softer incentives on the standard task, as well as the noisy one. The complementarity between the tasks (extreme in the version used here, but this is not necessary) dictates that as $e_N$ becomes more expensive for $\mathcal{P}$ because of the noise, he will choose to induce lower values of $e_S$ as well. Another way to check this is that the equilibrium values of both $w_S$ and $w_N$ decrease as $p$ goes down. When $p \leq 3/5$, $\sigma_N$ is not informative enough. In this case $e_S = e_N = w_{S1} = w_{N1} = 0$. This has been interpreted as no contract being signed. The no contract outcome obtains even though an informative contractible signal for both tasks is available.

$^3$See also Holmström and Milgrom (1994)
4. Informed Principal

Myerson (1983), Maskin and Tirole (1990) and Maskin and Tirole (1992) examine the case in which \( P \) has private information, creating a potential signaling role for the contract offer. Despite the intricacies involved, the simple benchmark model above can be adapted again to illustrate some of the key points.\(^4\)

There are two types of principal \( P_H \) and \( P_L \). \( P \) is of type \( H \) with probability \( \phi = 18/29 \) and of type \( L \) with probability \( 1 - \phi = 11/29 \). The principal’s type his private information. If \( P \) is of type \( H \), \( A \)'s outside option is \( k = 9/32 \), while if \( P \) is of type \( L \) then \( A \)'s outside option is 0, as in the benchmark above. Hence, if \( P_H \) and \( P_L \) separate in equilibrium, there are two \( IRs \) for \( A \), while if pooling obtains \( A \)'s expected outside option is \( \phi k = 81/464 \), and he faces a single \( IR \). \( A \)'s \( LCs \) are as in the benchmark above.

First \( P \) learns his type. Then he offers a contract to \( A \), which may take the form of a menu (wages contingent on output and \( P \)'s type). At this point \( A \) updates his beliefs about \( P \)'s type and then decides whether to accept or reject.\(^5\) After a contract is signed \( P \) tells \( A \) which part of the menu applies in his case (if the contract is in fact a menu). Finally, \( A \) chooses effort, output is realized and payoffs are obtained.

There is a single task requiring effort which stochastically produces output as in the benchmark model. Output is contractible. \( P \)'s payoffs and \( IR \) are also as above. \( A \)'s payoff is also as in the benchmark above, except that he takes expectations using his beliefs.

In a separating equilibrium \( P_H \) and \( P_L \) offer two distinct pairs of output-contingent wages: \((w_{H1}, w_{H0})\) and \((w_{L1}, w_{L0})\) respectively. \( A \)'s ICs dictate that after being offered \((w_{H1}, w_{H0})\) effort is \( e_H = w_{H1} - w_{H0} \), while after being offered \((w_{L1}, w_{L0})\) effort is \( e_L = w_{L1} - w_{L0} \).

Separation requires that neither \( P_H \) nor \( P_L \) has an incentive to offer the other type’s wage pair. Since \( P \)'s private information does not enter directly his payoff, this can only be true if the expected profits for the two types of principals, \( \Pi_H \) and \( \Pi_L \) are the same. This is the truth telling (henceforth TC) constraint, which, using IC, since \( w_{H0} \) can be shown to be 0, reads \( \Pi_H = e_H(1 - e_H) = e_L(1 - e_L) - w_{L0} = \Pi_L \). Since \( k = 9/32 \), one of the two \( IRs \) for the agent does bind. Using IC this yields \( e_H = w_{H1} = 3/4 \). Using TC, this implies \( e_L = 1/2, w_{L0} = 1/16 \) and \( w_{L1} = 9/16 \). With these values \( \Pi_H = \Pi_L = 3/16 \).

With informed principals, the literature highlights the possibility of pooling equilibria, in which the contract is a menu. Both \( P_H \) and \( P_L \) offer a menu \((w_{H1}^M, w_{H0}^M, w_{L1}^M, w_{L0}^M)\), which \( A \) has to accept or reject based on his expected \( IR \). After a contract is signed, \( P \) tells \( A \) which pair of

---

\(^4\)The computations below all pertain to the case of “Common Values” analyzed in Maskin and Tirole (1992).

\(^5\)As in any signaling game, the issue of off-the-equilibrium-path beliefs arises. The simplest way to deal with this issue is to assume that \( A \)'s beliefs after observing an “unexpected” offer are that \( P \) is of type \( H \) with probability 1. This is implicitly assumed in all computations below.
output-contingent wages applies. The TC constraint still applies, since both \( \mathcal{P}_H \) and \( \mathcal{P}_L \) have to be willing to indicate to \( \mathcal{A} \) the appropriate wage pair. In fact, using \( IC \) and \( w_{H0}^M = 0 \), \( IC \) still reads \( \Pi_H^M = e_H^M(1 - e_H^M) = e_L^M(1 - e_L^M) - w_{L0}^M = \Pi_L^M \). Using the single binding expected IR and the ICs, which are unchanged, yields \( (18/58)(e_H^M)^2 + (11/29)((e_L^M)^2 + w_{L0}^M) = 81/464 \). Using the TC constraint this gives \( e_H = w_{H1} = 5/8 \), \( e_L = 1/2 \), \( w_{L0} = 1/64 \) and \( w_{L1} = 33/64 \). With these values \( \Pi_H = \Pi_L = 15/64 \). Thus both types of \( \mathcal{P} \) enjoy strictly higher profits than under separation. Pooling relaxes \( \mathcal{A} \)'s IR which binds in expectation. \( \mathcal{P}_H \) can lower \( w_{H1} \) which increases \( \Pi_H^M \) relative to the separation case. The increased profit for \( \mathcal{P}_H \) affects \( \mathcal{P}_L \) via the TC constraint. \( \mathcal{P}_L \) lowers both output-contingent wages to satisfy the TC constraint, which in turn increases \( \Pi_L^M \) to keep it in line with \( \Pi_H^M \).

5. Intertemporal Incentives

Holmström and Milgrom (1987) analyze the case of a relationship between \( \mathcal{P} \) and \( \mathcal{A} \) that extends over time. Some of the main insights can be gained in the following simple set-up.

There are two time periods – the first denoted \( F \) and the second denoted \( S \). \( \mathcal{A} \) chooses an effort in \([0, 1]\) in both periods. Output can be either 1 or 0, and output draws are independent across the two periods. The first period effort is denoted \( e_F \). The second period effort if output is 1 in the first period is \( e_{1S} \), while the second period effort if output in the first period is 0 is \( e_{0S} \). The probability that output is 1 is \( \sqrt{e_F} \) in the first period, and \( \sqrt{e_{1S}} \) (with \( i \in \{0, 1\} \)) in the second period.

\( \mathcal{A} \) is paid at the end of the two periods, as a function of observed output in the two periods. The wage paid if output is \( i \in \{0, 1\} \) in period \( F \) and \( j \in \{0, 1\} \) in period \( S \) is denoted \( w_{ij} \).

Neither \( \mathcal{P} \) nor \( \mathcal{A} \) discount the future. While \( \mathcal{P} \) is risk-neutral, \( \mathcal{A} \) is risk-averse with an exponential utility with a constant absolute risk-aversion coefficient equal to 1/2. His effort in the two periods is perfectly substitutable. Given a wage scheme \( w_{ij} \) and effort levels \( e_F \) and \( e_{1S} \) his expected utility is

\[
\Pi^A = -\sqrt{e_F} \left[ \sqrt{e_{1S}} \exp\left\{ -\frac{1}{2}(w_{11} - e_F - e_{1S}) \right\} + (1 - \sqrt{e_{1S}}) \exp\left\{ -\frac{1}{2}(w_{10} - e_F - e_{1S}) \right\} \right]
- (1 - \sqrt{e_F}) \left[ \sqrt{e_{0S}} \exp\left\{ -\frac{1}{2}(w_{01} - e_F - e_{0S}) \right\} + (1 - \sqrt{e_{0S}}) \exp\left\{ -\frac{1}{2}(w_{00} - e_F - e_{0S}) \right\} \right]
\]

while \( \mathcal{P} \)'s expected payoff is

\[
\Pi^P = \sqrt{e_F} \left[ \sqrt{e_{1S}}(2 - w_{11}) + (1 - \sqrt{e_{1S}})(1 - w_{10}) \right]
+ (1 - \sqrt{e_F}) \left[ \sqrt{e_{0S}}(1 - w_{01}) + (1 - \sqrt{e_{0S}})(-w_{00}) \right]
\]
The optimal incentive scheme is found by maximizing $\Pi^P$ subject to $IR$ constraints imposing that $\Pi^A \geq -1$ and $\Pi^P \geq 0$ and subject to the $IC$ constraints which now impose that $e_F$, $e_{0S}$ and $e_{1S}$ should jointly maximize $\Pi^A$ given the incentive scheme $w_{ij}$.

The $IR$ constraint is binding for $A$, while it is not binding for $P$. The $IC$ constraint can be subsumed in the first order conditions obtained by differentiating $\Pi^A$ with respect to $e_F$ and $e_{iS}$ and setting these equal to 0 which are sufficient for a maximum.\footnote{These levels of reservation payoffs can be taken to be a normalization for $P$ and an assumption that $A$ can earn a certain payoff of 0 elsewhere, yielding a utility level of $-1$.}

To characterize the optimal incentive scheme for the two-period problem it is useful to first consider the second period ($S$) sub-problem after output $i \in \{0, 1\}$ has been realized in the first period ($F$). These problems are obtained considering (continuation) payoffs for $A$ and $P$ given by the relevant square bracket term of $\Pi^A$ and $\Pi^P$ above, and with an $IR$ constraint for $A$ given by his utility level (contingent on output in $F$) in the solution to the two-period problem, after factoring out the common term $\exp\{e_F/2\}$.

Using these binding $IR$ constraints and the first-order $IC$ constraints it can be seen that the difference $(w_{1i} - w_{i0}) > 0$ is independent of $i$ — the second-period incentive premium $\Delta_S = (w_{1i} - w_{i0})$ does not depend on first-period output. Hence, using the first-order $IC$ constraints it is also the case that $e_{0S} = e_{1S} = e_S \in (0, 1)$. $A$’s $IR$ constraints in each period $S$ sub-problem determines $w_{i0}$.

The period $S$ sub-problems can then be plugged into the two-period problem. Viewed from period $F$ we can think of $P$ as offering $A$ two certainty equivalent wages $c_i$ for each period $F$ output. Notice that we can write $c_i = \tilde{w}_i - \pi_i$ where $\tilde{w}_i$ is the expected period $S$ wage when the realized period $F$ output is $i$ and $\pi_i$ is the associated risk-premium. Since $(w_{1i} - w_{i0}) = \Delta_S$ is independent of $i$, and $A$’s utility exhibits constant absolute risk-aversion we then get $\pi_0 = \pi_1 = \pi$. Hence factoring out the common term $\exp\{\pi/2\}$ from $A$ utility, the period $F$ problem can be seen as having the same form as the two period $S$ sub-problems with a different $IR$ constraint for $A$. Hence, as before, the difference $\Delta_F = (\tilde{w}_1 - \tilde{w}_0)$ does not depend on $A$’s reservation utility and in fact $\Delta_F = \Delta_S = \Delta$. For the same reason $e_F = e_S = e$.

Using $\Delta_F = \Delta_S = \Delta$ and $e_F = e_S = e$ we then get that the optimal incentive scheme is linear in output in the sense that $w_{01} = w_{10} = w_{10} + \Delta$ and $w_{11} = w_{00} + 2\Delta$. Given $w_{10}$, the wage increases by a fixed amount $\Delta$ for each unit of realized output over the two periods.

\footnote{This way to proceed is known in the literature as taking the first-order approach. In the more general case considered for instance by Holmström and Milgrom (1987) this is not viable. In the simple case considered here, the first-order approach works because we are assuming that the exponent of effort variables (1/2 in this case) plus $A$’s constant absolute risk-aversion coefficient (also 1/2 in this case) sum to 1.

Even in single-period agency models, whether the first-order approach is valid or not is an intricate question first uncovered by Mirrlees (1975). Subsequent contributions on this topic can be found in Grossman and Hart (1983), Rogerson (1985) and Jewitt (1988).}
In the simple model we have used here output is either 1 or 0. The linearity result holds in the same model (with an arbitrary finite number of periods) when there are \( N \) possible output realizations each period. In this case the incentive scheme is linear in accounts — in essence linear in a vector of variables that count the number of realizations of each possible output level.

Hellwig and Schmidt (2002) clarify that linearity in accounts need not imply linearity in aggregate output, and in fact some additional assumptions are needed for the latter to hold. They show that if \( A \) can destroy output unnoticed, and \( P \) only observes aggregate output at the end of the last period, then the (approximately) optimal incentive scheme is indeed linear in aggregate output.

Both Holmström and Milgrom (1987) and Hellwig and Schmidt (2002) are principally concerned with a continuous-time model in which \( A \) controls the drift of a (multi-dimensional) Brownian motion process that represents output. The continuous time version of the problem yields elegant closed-form solutions that confirm the linearity result. Hellwig and Schmidt (2002) analyze in detail the status of the continuous time model as the limit of discrete time models.

The linearity of incentive schemes is of great interest in applications because of the prominence in practice of linear (or approximately linear) incentive schemes. In all known theoretical settings, linear optimal incentive schemes rely on exponential utility functions for both \( A \) and \( P \), whenever he is not risk-neutral. Stochastically independent periods also play a crucial role.

Finally, the tight linear characterizations of intertemporal incentive schemes also rely on \( P \)’s ability to commit in advance to an incentive scheme, and on \( A \)’s ability to commit not to quit before the end. The question of whether a full-commitment long-term contract can be implemented via a sequence of short-term contracts has been analyzed in a general context by Malcomson and Spinnewyn (1988), Fudenberg, Holmström, and Milgrom (1990) and Rey and Salanié (1990). A common thread of this literature is that \( P \)’s ability to monitor \( A \)’s savings decisions plays a key role in the possibility of short-term implementation of long-term contracts.

### 6. Recent Developments

Since its inception the literature on agency problems and applications has grown dramatically, influencing many areas of Economics ranging from Development to Finance. Recent developments in the actual analytical framework relax some of the basic assumptions of the canonical model.

Eliaz and Spiegler (2006) and O’Donoghue and Rabin (2005) focus on the underlying behavioral assumptions. The first paper tackles an environment in which agents may differ in their cognitive abilities which generates dynamically inconsistent behavior. The second paper is concerned with

---

8Agency Theory has found a prominent place in many graduate and undergraduate programs in Economics. Recent texts that provide a comprehensive treatment of the field include Salanié (2000), Laffont and Martimort (2002) and Bolton and Dewatripont (2005).
the effect of present-bias in the agent’s preferences on the optimal incentive scheme. In both cases the optimal incentive scheme becomes more realistically “sensitive to detail” than in the standard case.

Besley and Ghatak (2006) focus on the case of motivated agents in the provision of a public good. Motivated agents do not always regard effort as a cost. This has important effects on incentive design, which in turn sheds light on the nature of non-profit organizations.

References


