1. How does this fit the stag hunt example? (*Hint:* The receiver is the benevolent chief of the village, and $y = 1$ if and only if he has decided to send everybody hunting.)

2. We saw in section 4.3.1 that it is reasonable to look for equilibria in which each sender i announces "yes" if, and only if, $m_i \leq m$ and "no" otherwise. Show that in any such equilibrium, there cannot be a switch from $y = 1$ to $y = 0$ if one villager changes his "no" to a "yes."

3. Show that m must be equal to 1.

4. (*Slightly more difficult*) Compute the equilibrium probability that all go hunting, and show that it converges to 1 as ε becomes arbitrarily small.

References

Aghion, P., and J. Tirole. 1997. Formal and real authority in organizations. *Journal of Political Economy* 105:1–29.

Akerlof, G. 1970. The market for lemons: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 89:488–500.

Benabou, R., and G. Laroque. 1992. Using privileged information to manipulate markets: Insiders, gurus, and credibility. *Quarterly Journal of Economics* 107:921–58.

Cho, I.-K., and D. Kreps. 1987. Signaling games and stable equilibria. *Quarterly Journal of Economics* 102:179–221.

Crawford, V., and J. Sobel. 1982. Strategic information transmission. *Econometrica* 50:1431–51.

Leland, H., and D. Pyle. 1977. Asymmetries, financial structure, and financial intermediation. *Journal of Finance* 32:371–87.

Mas-Colell, A., M. Whinston, and J. Green. 1995. *Microeconomic Theory*. Oxford: Oxford University Press.

Maskin, E., and J. Tirole. 1992. The principal–agent relationship with an informed principal. II: Common values. *Econometrica* 60:1–42.

Spence, M. 1973. Job market signaling. *Quarterly Journal of Economics* 87:355–74.

5 Moral Hazard

Well, then, says I, what's the use you learning to do right when it's troublesome to do right and ain't no trouble to do wrong, and the wages is just the same? I was stuck. I couldn't answer that. So I reckoned I wouldn't bother no more about it, but afterwards always do whichever come handiest at the time.

—**Mark Twain,** *Adventures of Huckleberry Finn.*[1]

We speak of *moral hazard* when

- the Agent takes a decision ("action") that affects his utility and that of the Principal;
- the Principal only observes the "outcome," an imperfect signal of the action taken;
- the action the Agent would choose spontaneously is not Pareto-optimal.

Because the action is unobservable, the Principal cannot force the Agent to choose an action that is Pareto-optimal. He can only influence the choice of action by the Agent by conditioning the Agent's utility to the only variable that is observable: the outcome. This in turn can only be done by giving the Agent a transfer that depends on the outcome.

1. Quoted by Holmstrom-Milgrom (1987).

Examples of moral hazard abound, and it is difficult to imagine an economic relationship that is not contaminated by this problem.[2] If a perfect relationship could exist, the Principal would be able to observe all the decision variables of the Agent that relate to his utility; this would be extremely costly in terms of supervisory measures.

Moral hazard is present everywhere within firms, since employers rarely can control all decisions of their employees. The term *effort* is often used to designate the employee inputs that are not directly observable, the employer can only base wages on production or some other observable variable that induces employees not to shirk. This term effort is confusing in that it suggests that moral hazard in firms consists only in employees avoiding work. However, moral hazard exists as soon as the objectives of the parties differ. A good example is the relationships between shareholders and managers. Because the managers are autonomous agents, they will have objectives that are not necessarily the same as those of the shareholders (who above all want the firm's value to be maximized).

In the area of property insurance, the moral hazard is due to an insurer not being able to observe the precautions against theft, fire, and so forth, of the insured despite the positive effects of such effort on the insurer's profits.

In service activities, moral hazard is present where the effort of the service provider bears on the outcome of a task. Simple examples include the relationship between a car-owner and his mechanic, or between a patient and his doctor.

Last, in the economics of development, moral hazard is often studied to describe the relationships between landowners and their farmers. In sharecropping, for example, agreements stipulate that the harvest will be shared between both parties, thus making it important for the landlord to get the farmer to put in effort.

2. The moral hazard model actually is often called the "agency problem" and identified with the Principal–Agent model.

The first-best situation is therefore defined by the situation where the Principal can observe every action of the Agent. Then he can rcommend that the Agent choose the most efficient action,[3] and the wages that provide for optimal risk sharing. It is often assumed that in these models the Principal is risk-neutral; for instance, the Principal faces many independent risks and thus can diversify the risks associated to his relationship with the Agent.[4] In contrast, the Agent normally exhibits risk-aversion (it is more difficult for him to diversify his risks). Optimal risk sharing then requires that the Principal perfectly ensure the Agent by paying him a constant wage and by bearing all risks involved in their common activity.

In the second-best situation the Principal can only observe a variable correlated with the Agent's action: the outcome. If the Principal is risk-neutral, the first-best optimum consists in giving the Agent a constant wage. In second-best circumstances this will tempt the Agent to choose selfishly the action that is the least costly for him, and in general, this is not optimal.[5] Solving the moral hazard problem thus implies that the Principal offers the Agent a contract with trade-offs between risk sharing and incentives:

- Risk sharing so that the Agent's wage do not depend too heavily on the outcome.
- Incentives so that the Principal can base the Agent's wage on the outcome.

Now, when the Agent is risk-neutral, this trade-off is nonexistent. The Agent does not mind bearing all the risk, so the issue of risk-sharing is irrelevant. We sometimes say in that case that the moral hazard problem is solved by "selling the firm to the Agent." However, this case has little practical interest.

3. Or, equivalently, the Principal can fine the Agent if he does not choose the efficient action.
4. This is by no means always the most natural assumption, as the patient–doctor relationship shows. However, it is not crucial to the analysis.
5. This is the meaning of the *Huckleberry Finn* quotation that opens this chapter.

5.1 A Simple Example

We start with the simplest framework: a two action, two outcome model. The Agent can choose between working, $a = 1$, and not working, $a = 0$. The cost of action a is normalized to a so that the Agent's utility, if he gets wage w and chooses action a, is $u(w) - a$, where u is strictly concave. The Principal can only observe whether the Agent succeeds or fails at his task. If the Agent works, his probability of succeeding is P and the Principal gets a payoff x_S. If he does not work, the probability of success falls to $p < P$, and the Principal's payoff is $x_F < x_S$.

In the more interesting case the Principal must induce the Agent to work. Then he has to give the Agent wages w_S (in case of success) and w_F (in case of failure) such that the Agent's effort is rewarded:

$$Pu(w_S) + (1 - P)u(w_F) - 1 \geq pu(w_S) + (1 - p)u(w_F)$$

so the incentive constraint is

$$(P - p)(u(w_S) - u(w_F)) \geq 1$$

Because the Principal must (obviously) pay a higher wage when the Agent works, the difference $(w_S - w_F)$ increases as P gets closer to p. As this occurs, it becomes difficult to distinguish a worker from a nonworker. Then we say that the incentive to work must become more *high powered.*

We must also take into account an individual rationality constraint. By this we mean that the Agent must find it worthwhile to work rather than to quit and get his outside option $\underline{U}$. This gives

$$Pu(w_S) + (1 - P)u(w_F) - 1 \geq \underline{U}$$

This inequality must clearly be an equality. Otherwise, the Principal can decrease both $u(w_S)$ and $u(w_F)$ by the same small transfer ε, which would not affect the incentive constraint and would increase his own utility, since (assuming he is risk-neutral) this is

$$P(x_S - w_S) + (1 - P)(x_F - w_F)$$

Proving that the incentive constraint is an equality is slightly more involved. If it were a strict inequality, we could subtract $(1 - P)\varepsilon/u'(w_S)$ from w_S and add $P\varepsilon/u'(w_F)$ to w_F. The incentive constraint would still hold for ε small. By construction, $u(w_S)$ would decrease by $(1 - P)\varepsilon$ and $u(w_F)$ would increase by $P\varepsilon$ so that the individual rationality constraint would still be satisfied. Moreover the wage bill $Pw_S + (1 - P)w_F$ of the Principal would decrease by $P(1 - P)\varepsilon(1/u'(w_S) - 1/u'(w_F))$, which is positive because $w_F < w_S$ and u is strictly concave.[6]

Because both inequalities are linear equalities in $(u(w_F), u(w_S))$ and we have just two unknowns, we can easily solve for $u(w_S)$ and $u(w_F)$. This gives

$$\begin{cases} u(w_F) = \underline{U} - \dfrac{p}{P - p} \\ u(w_S) = \underline{U} + \dfrac{1 - p}{P - p} \end{cases}$$

from which we can proceed to compute the Principal's expected utility

$$W_1 = P(x_S - w_S) + (1 - P)(x_F - w_F)$$

However, this is a very special case. We only relied on the maximization of W_1 to prove that both constraints are binding at the optimum.

It might well be that the Principal finds it too costly to get the Agent to work and decides to let him shirk instead. In this case he

6. More diagram-oriented readers can also easily see this by drawing a curve in the $(u(w_F), u(w_S))$ plane.

will give the Agent a constant wage $w_S = w_F = w$ such that $u(w) = \underline{U}$, and he will get an expected utility

$$W_0 = px_S + (1 - p)x_F - w$$

The difference between W_0 and W_1 can then be rewritten as

$$W_1 - W_0 = (P - p)(x_S - x_F) + w - Pw_S - (1 - P)w_F$$

Since the wages do not depend on x_S and x_F, it appears that if success is much more attractive than failure for the Principal ($x_S - x_F$ is high), he will choose to get the Agent to work. (The reader is asked in exercise 5.1 to prove that then $x_S - w_S > x_F - w_F$ at the optimum, with the surplus from success shared between the Agent and the Principal.)

5.2 The Standard Model

We consider here the standard model in a discrete version. The Agent can choose between n possible actions: $a_1, \ldots, a_n$. These actions produce one among m outcomes, which we denote $x_1, \ldots, x_m$.

The outcome a priori is a signal that brings information on the action the Agent chooses. To simplify matters, we identify it as surplus from the Principal–Agent relationship.[7] (We will return to this assumption in section 5.3.4.)

The stochastic relationship between the chosen action and the outcome is often called a "technology." The idea here is that when the Agent chooses action a_i, the Principal observes outcome x_j with a probability p_{ij} that is positive.[8] Because the only variable that is publicly observed is the outcome, contracts must take the form of a wage that depends on the outcome. If the Principal observes the outcome x_j, he will pay the Agent a wage w_j and keep $x_j - w_j$ for himself.

A general specification for the Agent's von Neumann–Morgenstern utility function would be $u(w, a)$. However, the choice of action would then affect the agent's preferences toward risk, which would complicate the analysis.[9] Therefore we will assume that the Agent's utility is separable in income and action. Moreover it is always possible to renormalize the actions so that their marginal cost is constant.

Thus in the standard model we take the Agent's utility function to be

$$u(w) - a$$

where u is increasing and concave. We can assume that the Principal is risk-neutral, as done in most of the literature. The Agent's von Neumann–Morgenstern utility function then is

$$x - w$$

5.2.1 The Agent's Program

When the Principal offers the Agent a contract w_j, the Agent chooses his action by solving the following program:

$$\max_{i=1,\ldots,n} \left(\sum_{j=1}^{m} p_{ij} u(w_j) - a_i \right)$$

If the Agent chooses a_i, then the $(n - 1)$ incentive constraints

$$\sum_{j=1}^{m} p_{ij} u(w_j) - a_i \geq \sum_{j=1}^{m} p_{kj} u(w_j) - a_k \quad (IC_k)$$

must hold for $k = 1, \ldots, n$ and $k \neq i$.

7. For instance, in an employer–employee relationship, a will be the effort and x the resulting production or profit.

8. If some of the probabilities p_{ij} were zero, the Principal could use this information to exclude some actions. Suppose that action a_i is the first-best optimal action and that $p_{ij} = 0$ for some j. The Principal then can fine the Agent heavily when the outcome is x_j, since the fact that he observes x_j signals that the Agent did not choose the optimum action a_i. This type of strategy will even allow the Principal to implement the first-best: if moreover $p_{kj} > 0$ for all $k \neq i$, then the choice of any a_k other than a_i will expose the Agent to a large fine, thus effectively deterring him from deviating. This was noted early on by Mirrlees (1975, published 1999); it is the reason why I exclude this case.

9. Then it may be optimal for the Principal to give higher wages if it reduces the Agent's disutility of effort, so that the individual rationality constraint may not be binding at the optimum.