Review of Economic Studies (1999) **66**, 3–21 0034-6527/99/00010003$02.00

The Theory of Moral Hazard and Unobservable Behaviour: Part I

J. A. MIRRLEES
Trinity College, Cambridge

This paper was completed in October 1975 *but never previously published* (*Eds.*)

Principal-agent models are studied, in which outcomes conditional on the agent's action are uncertain, and the agent's behaviour therefore unobservable. For a model with bounded agent's utility, conditions are given under which the first-best equilibrium can be approximated arbitrarily closely by contracts relating payment to observable outcomes. For general models, it is shown that the solution may not always be obtained by using the agent's first-order conditions as constraint. General conditions of Lagrangean type are given for problems in which contracts are finite-dimensional.

1. THE MORAL HAZARD SITUATION

Appreciation that unobservable behaviour by insured persons, and others participating in contingent contracts, is an interesting and important subject for economists, is due to Arrow (1970*b*) and Pauly (1968). Neither author showed how to model these situations formally, nor suggested any worthwhile propositions. They agreed that insured persons who fully exploit their contracts, expressed in terms of observable behaviour, thereby reduce the efficiency of the economy. Pauly adopted the startling position that such "rational economic behaviour" cannot be morally perfidious; Arrow, more reasonably, emphasized its disadvantages. But both were wrong: there is a wide class of cases in which there is no significant loss of efficiency as a result of self-interested unobservable behaviour. This will be demonstrated in Section 3 below.

The first formal model of moral hazard, that I am aware of, was that of Zeckhauser (1970), who discussed the choice of an insurance policy for which payments are proportional to medical expenditures, with individuals choosing their medical expenditures to suit themselves. Solutions were computed using first-order conditions. Spence and Zeckhauser (1971) placed such problems in the general setting of behaviour under uncertainty subject to deliberately chosen contingent contracts. Their paper is largely taxonomic, and does no more than derive first-order conditions for the general (non-linear) case. These conditions are not rigorously justified, and are in any case incomplete. A model of precisely this kind was used in Mirrlees (1972) to discuss the relationship of taxation to family size: in that paper some qualitative results were obtained, about the form of the optimal contingent "contract" (in this case, a tax system). Again, the argument relies on first-order conditions; and in this paper a complete set is used, in the sense that they can fully determine the optimum. None of these papers attempts to justify the first-order conditions rigorously.

It is already clear in the Spence–Zeckhauser formulation that the situation first considered for the case of moral hazard in insurance is in fact a very general one, arising whenever behaviour is unobservable, but its consequences are observable. The same issues arise in the analysis of share-cropping (Cheung (1969) and Stiglitz (1974)); in an adequate study of capital markets, credit, and lending; in the theory of agency (Ross (1973*a*, *b*)); and the theory of incentive systems and pay structures (Stiglitz (1975) and Mirrlees

(1976)). The problem proved, therefore, to be a serious stumbling block in the formulation of a satisfactory general equilibrium theory under uncertainty (Radner (1968)).

We shall study models in which there is uncertainty about the outcome of people's actions, the actions being themselves unobservable though the outcomes are observable. Contracts in terms of outcomes can be entered into. These might be contracts between pairs of agents, but the most interesting ones are multilateral, covering large numbers of agents. This multilateral aspect normally appears in the form of insurance, private or public. A major issue is the identification of optimal insurance schemes, relative to the assumption that promises about unobservable behaviour are not made, nor do moral considerations govern behaviour. (Hence, as John Flemming has pointed out, it is odd that the problem of self-interested unobservable behaviour has come to be called "moral hazard.") It is also interesting to estimate the social gains available from moral behaviour, *i.e.* the choice of unobservable actions in the social rather than the private interest.

Some of the main results of this paper have previously been sketched, for a special case, in Mirrlees (1974). The arguments in that paper are only heuristic; the issue of how general the results are is not discussed; and the significance and usefulness of the results and methods only touched upon. The question of justifying first-order conditions that have been obtained heuristically has been mentioned several times in this introduction. It turns out to be neither an easy nor an insignificant question. It is perfectly possible that the first-order conditions used in Mirrlees (1974) and earlier papers will fail to find the optimum in certain cases. Furthermore these cases are not exceptional (like those for which Lagrange multipliers or Kuhn–Tucker conditions fail to work). The difficulty alluded to is quite different from those that have arisen in other maximization problems, and seem to be much harder to deal with. All this is explained in Section 4 of the present paper.

The paper is devoted to a rigorous analysis of unobservable behaviour in a fairly general form. Nevertheless, it should not be found highly technical: the mathematical arguments, though sometimes lengthy, are never very difficult, and remain quite close to intuitive essentials. The basic model is formulated in Section 2. There are then two main cases to be considered, those of unbounded and bounded utility. A theorem about the first case is proved in Section 3: this result is probably the most striking in the paper. The calculation of solutions to cases with bounded utility is discussed, with examples, in Section 5,[1] and in Section 6 an interesting special case is analysed in some detail. Sections 7 and 8 extend the analysis to circumstances in which differences among individuals have to be taken into account. Section 8 considers some further extensions, including more sophisticated welfare functions. Section 9 briefly proposes some implications and applications.

2. IDENTICAL AGENTS WITH EQUIVALENT INFORMATION

Although, as the introduction has emphasized, there are many situations in which behaviour is affected by rewards of uncertain application, it is nevertheless stimulating to have a particular situation in mind when analysing the abstract problem. Since moral hazard was first considered in the context of insurance, let us focus on individuals indulging in an activity subject to a risk of loss through accidents. We enquire what insurance arrangements may be best for a large group of such individuals with independent accident prospects. In the first model to be considered, these individuals are identical in preferences and opportunities, correctly assess probabilities, and make choices so as to maximize expected utility: it is not unreasonable, therefore, to seek insurance arrangements that maximize expected utility. Such an arrangement will be termed optimal.

1. Editors note. These sections are all contained in the (unavailable) Part II of this paper.

The following notation is used, with reference to a typical individual:

x = net compensation, best taken as income after allowing for insurance premia and compensation;
y = losses arising from accidents;
z = care, *i.e.* the effort made to reduce accident losses;
w = the state of nature, calibrated in such a way as to be uniformly distributed over the interval [0, 1].

x, y, z are totals over some period of time, which ought to be the whole period of participation in the activity considered. We do not, at this stage, allow for variations in care over time, or attach any importance to the timing of losses and compensation payments.

The individual wishes to maximize

$$\mathrm{Eu}\,(x, y, z) = \int_0^1 u(x, y, z)dw, \tag{1}$$

subject to insurance arrangements, described by

$$x = a(y), \tag{2}$$

and loss possibilities, described by

$$y = b(z, w). \tag{3}$$

x, y, z are non-negative numbers. In the terminology of Spence and Zeckhauser, z, and also the function a, are chosen before w is known.

u *is a twice continuously differentiable strictly concave function of* (x, y, z), *increasing in* x *and decreasing in* y, (4)

b *is a twice continuously differentiable decreasing function of* z, (5)

a *can be any integrable function from the non-negative real line to itself.* (6)

The choice of compensation arrangements is subject to feasibility, which is expressed by

$$Ex \leqq g(Ey). \tag{7}$$

The case of an unsubsidized mutual insurance scheme is $g \equiv 0$. It will be assumed that g is a decreasing function. (7) makes a large numbers assumption, that, despite uncertainty, total income and total losses are (multiples of) expected income and losses for an individual. The importance of this assumption will be explored below.

In formal terms, the problem is to choose the function a so as to maximize (if possible) Eu, subject to two constraints, namely the constraint that

$$z \text{ maximizes } \mathrm{Eu}\,(a(y), y, z) = \int_0^1 u(a(b(z, w)), b(z, w), z)dw, \tag{8}$$

and the feasibility constraint (7) for this value of z. This is a double maximization problem of the type

$$\max_a \left\{ \max_z U(a, z) : (a, z) \in S \text{ with } z \text{ maximizing } U \right\}, \tag{9}$$

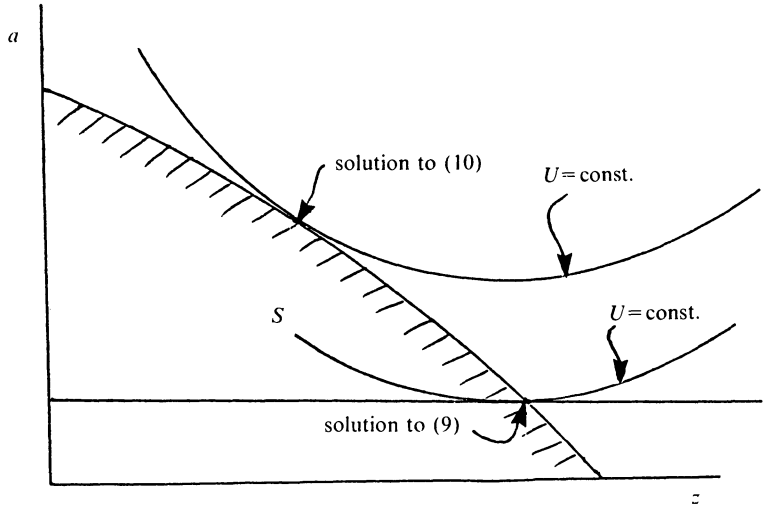

FIGURE 1

where S is a feasibility set. It should be carefully noted that (9) is not in general the same as

$$\max_{a}\left\{\max_{z}\left[U(a,z):(a,z)\in S\right]\right\}, \tag{10}$$

although the maximum in (9) cannot be greater than the maximum in (10) if the two maxima are attained. The point is that, when $\bar{a}$, $\bar{z}$ are such as to maximize U over S, there is no reason why $\bar{z}$ should maximize $U(\bar{a}, z)$ when z is unrestricted. This is exemplified in Figure 1, where a as well as z is taken to be a scalar (a parameter in the compensation function, perhaps).

The second problem (10) describes the choice of insurance arrangements when the behaviour of individuals can be determined independently of the choice of a, for example because it is possible to rely on, or costlessly police, prior agreements by all parties to the insurance scheme about the care they will take. A moral hazard problem is, by definition, one where the agents have to be assumed to choose z to suit themselves individually after a has been chosen. Nevertheless, it is interesting to compare the outcome of problem (9), based on independent individual action, with the outcome of problem (10), where behaviour must be "centrally" determined. If (9) has a solution, that solution will be called the optimum; while a solution for (10) will be called the full-control optimum.

Returning to the main problem, we note a reformulation which is rather useful. The idea is to eliminate w by considering the distribution of y conditional upon z. The distribution function for y, given z, being defined as $F(y, z)$, we have $w = F(y, z)$, and F is therefore given by (3)

$$y = b(z, F(y, z)). \tag{11}$$

The density function $f(y, z) = F_y$ is deduced from (11)

$$f(y, z) = 1/b_w(z, F(y, z)). \tag{12}$$

From this point of view, the problem is to

$$\left.\begin{aligned} &\operatorname*{Max}_{a} \int u(a(y), y, z) f(y, z)\,dy, \\ &\text{subject to } z \max \int u(a(y), y, z) f(y, z)\,dy, \\ &\text{and} \quad \int a(y) f(y, z)\,dy \leqq g\left(\int y f(y, z)\,dy\right). \end{aligned}\right\} \tag{13}$$

It should be noted that our assumption that b is a decreasing function of z implies that

$$F_z > 0. \tag{14}$$

To prove this, differentiate (11) with respect to both y and z

$$1 = b_w F_y = b_w f, \qquad 0 = b_z + b_w F_z.$$

Thus

$$F_z = -b_z f,$$

which is positive, by (5), as was claimed. Furthermore, since for all positive z,

$$\begin{aligned} &F(0, z) = 0 \quad \text{and} \quad \lim_{y \to \infty} F(y, z) = 1, \\ &F_z = 0 \quad \text{at } y = 0 \quad \text{and} \quad F_2 \to 0 \text{ as } y \to \infty. \end{aligned} \tag{15}$$

This formulation is somewhat simplified mathematically, and will be used in the next two sections, where the theory of the problem is developed. But it is not well adapted to discussing how increases in the information available affect the outcome. When we come, in Section 6, to examine that issue, we shall find the original formulation of the problem more appropriate.

3. AN UNPLEASANT THEOREM

Many different assumptions about u and f are of interest. In the present section, it is supposed that it is possible to adapt and enforce contracts that have, in some states of nature, arbitrarily bad consequences for the individual; and that, though all loss levels are possible and significantly affected by care, the probability of losses falls off fairly rapidly at high levels. To be more precise, the following assumptions are made:

$$\begin{aligned} &\text{For all positive } y \text{ and } z, \\ &u(x, y, z) \to -\infty \qquad (x \to 0); \end{aligned} \tag{16}$$

$$\begin{aligned} &\text{For all positive } z, \\ &f_z(y, z)/f(y, z) \to -\infty \qquad (y \to \infty). \end{aligned} \tag{17}$$

The first of these assumptions needs no further discussion at present: like the second, it is often inapplicable. Perhaps it is never strictly applicable, but it allows us to make an important point sharply. The second assumption is often applicable, for it is satisfied by

the following special cases, involving exponential and lognormal distributions:

$$f = ze^{-zy};\ f = (2\pi)^{-1/2}(z/y)\exp[-\tfrac{1}{2}(\log y + \log z)^2].$$

More generally, we have the following criterion, which justifies the verbal description given initially:

Lemma. *Suppose*

(i) $$\inf_y \frac{z}{y}\left(\frac{\partial y}{\partial z}\right)_w > 0 \qquad \textit{for all positive } z, \tag{18}$$

(ii) *for all positive* z,

$$\frac{1-F}{yf} = \frac{1}{yf(y,z)}\int_y^\infty f(y',z)dy' \to 0 \qquad (y \to \infty), \tag{19}$$

(iii) $\lim_{y\to\infty}(f_z/f)$ *exists.*

Then

$$f_z/f \to -\infty \quad (y \to \infty).$$

Proof. $(\partial y/\partial z)_w = -F_z/F_y$, since $w = F(y, z)$. Therefore, by (18) and (19)

$$\frac{z}{y}\frac{F_z}{F_y}\frac{yf}{1-F} \to -\infty \qquad (y \to \infty).$$

As $F_y = f$, this can be written

$$z\int_y^\infty f_z(y',z)dy' \Big/ \int_y^\infty f(y',z)dy' \to -\infty.$$

It follows that, since f_z/f tends to a limit, the limit is $-\infty$, as was to be proved. ||

The elasticity condition (18) is satisfied if, for example, the production function is $y = b(z, w) = b_1(z)b_2(w)$, *i.e.* if care can be so measured as to have a proportional effect on losses. (19) is satisfied if $f > 0$ for all y, but also f decreases with at least exponential rapidity.

The peculiar interest of these assumptions arises from the following result:

Theorem 1. *If*

(i) $$\lim_{x\to 0} u(x, y, z) = -\infty \qquad \textit{for all positive } y \textit{ and } z, \tag{16}$$

(ii) $$\lim_{y\to\infty} f_z/f = -\infty \qquad \textit{for all positive } z, \tag{17}$$

(iii) $$u_{xz} \lesseqgtr Qu_x \qquad \textit{for some number } Q, \textit{ and all } x, y, z, \tag{20}$$

(iv) $$u_{xy} \gtreqless 0. \tag{21}$$

Then

$$\sup_a \left\{ \int_0^\infty u(a(y), y, z) f(y, z) dy : z \text{ maximizes } \int_0^\infty u f dy, \int_0^\infty a f dy \leqq g\left(\int_0^\infty y f dy \right) \right\}$$

$$= \max_{a,z} \left\{ \int_0^\infty u(a(y), y, z) f(y, z) dy : \int_0^\infty a f dy \leqq g\left(\int_0^\infty y f dy \right) \right\} \quad (22)$$

provided the latter is attained with positive a. The supremum is not attained for any positive function a.

In words, it is possible to approximate arbitrarily closely to, but not to attain, the full-control optimum.

Remarks. It will emerge in the course of the proof that the full-control optimum can be actually attained only in a very special case. The nature of the approximately optimal policies recommended by the theorem will be discussed after the proof.

Assumption (20) is rather weak, yet stronger than is actually required: its merit is simplicity. Assumption (21) is very plausible for the insurance case: it is satisfied if, for example, u takes the form $u(x-y, z)$, allowing perfect compensation. When (21) does not hold, a theorem of the above kind can be proved, as we shall see. It should also be recollected that u is concave, $u_x > 0$, and $F_z > 0$.

Proof. Since a full-control optimum exists with positive income, which we denote by a_0, z_0, it satisfies, for some constant λ,

$$u_a(a_0(y), y, z_0) = \lambda, \quad (23)$$

$$\frac{\partial}{\partial z} \left\{ \int_0^\infty u(a_0(y), y, z) f(y, z) dy - \lambda \int_0^\infty a_0(y) f(y, z) dy + \lambda g\left(\int_0^\infty y f dy \right) \right\}$$

$$= 0 \text{ at } z = z_0. \quad (24)$$

(23) follows from the possibility of transferring income from one state of nature (output level) to another without affecting the feasibility constraint; and then (24) is obtained by considering variations in z along with changes in income sufficient to maintain the feasibility constraint.

Define $U_0(z) = \int_0^\infty u(a_0(y), y, z) f(y, z) dy$. (24) implies that

$$U_0'(z_0) = \lambda \int_0^\infty \left\{ a_0(y) - g'\left(\int y f dy \right) y \right\} f_z(y, z_0) dy$$

$$= -\lambda \int_0^\infty \{a_0'(y) - g'\} F_z(y, z_0) dy, \quad (25)$$

on integrating by parts. By (14), $F_z > 0$; $g' < 0$ by assumption; and $\lambda > 0$ by (23). Therefore $U_0'(z_0) < 0$ if $a_0'(y) \geqq 0$. Differentiating (23), we have

$$a_0'(y) = -u_{xy}/u_{xx} \geqq 0,$$

by assumption (iv) and concavity. Therefore (25) does imply that

$$U_0'(z_0) < 0. \quad (26)$$

This means that z_0 does not maximize U_0. Consequently no income function and care level which are together feasible and such that z maximizes U can actually maximize U subject to feasibility alone: the full-control optimum is not attainable.

To prove the theorem, we construct income functions a_M with the properties that

$$\int_0^\infty a_M(y)f(y, z_M)dy \leqq g\left(\int_0^\infty yf(y, z_M)dy\right), \tag{27}$$

where z_M is such that

$$z_M \text{ maximizes } U_M(z) = \int_0^\infty u(a_M(y), y, z)f(y, z)dy, \tag{28}$$

and further

$$\int_0^\infty u(a_M(y), y, z_M)f(y, z_M)dy \to \int_0^\infty u(a_0(y), y, z_0)f(y, z_0)dy$$
$$\text{as } M \to \infty. \tag{29}$$

It is tempting to choose a_M in such a way that $z_M = z_0$, by using the first-order condition for (28) (as was done in the heuristic argument presented in Pauly (1968)); but it is then hard to prove global maximization of expected utility. Here it is arranged that $z_M \geqq z_0$, and this is sufficient for our purposes.

Define

$$\left.\begin{aligned} a_M(y) &= a_0(y), \qquad (y \leqq M), \\ &= \delta a_0(y), \qquad (y > M), \end{aligned}\right\} \tag{30}$$

where δ, a number between 0 and 1, remains to be chosen, but is in any case independent of y. Since u is an increasing function of x,

$$\int_0^\infty u(a_M(y), y, z)fdy < \int_0^\infty u(a_0(y), y, z)fdy, \qquad \text{for all } z, \tag{31}$$

or more concisely,

$$U_M(z) < U_0(z).$$

By direct calculation, we have

$$\begin{aligned} \frac{d}{dz}\{U_0(z) - U_M(z)\} &= \int_0^\infty \int_{a_M}^{a_0} u_{xz}(x, y, z)dxfdy \\ &\quad + \int_0^\infty \{u(a_0, y, z) - u(a_M, y, z)\}f_z dy \\ &\leqq Q\int_0^\infty \{u(a_0, y, z) - u(a_M, y, z)\}fdy \\ &\quad + \int_M^\infty \{u(a_0, y, z) - u(a_M, y, z)\}f_z dy, \end{aligned} \tag{32}$$

by (20) and (30).

By assumption (ii), given any number K, there exists M such that

$$f_z < -(K+Q)f, \qquad (y \geqq M). \tag{33}$$

From (32), we then have for this value of M

$$\frac{d}{dz}\{U_0(z) - U_M(z)\} < -\int_M^\infty \{u(a_0, y, z) - u(a_M, y, z)\} f dy$$
$$= -K\{U_0(z) - U_M(z)\} \tag{34}$$

In particular, $U_0 - U_M$ is a decreasing (positive) function of z. Let us now choose δ (if possible) so that

$$K\{U_0(z_0) - U_M(z_0)\} = -\inf_{z \leqq z_0} U_0'(z) = A. \tag{35}$$

Since U_0' is continuous, the constant A on the right here is finite, and positive by (26). $U_0(z_0) - U_M(z_0) = \int_M^\infty \{u(a_0, y, z_0) - u(\delta a_0, y, z_0)\} f dy, z_0) dy$ is a continuous function of δ, equal to zero when $\delta = 1$, and, by assumption (i), tending to infinity as $\delta \to 0$. Therefore there is indeed a (unique) value of δ satisfying (35).

Then, since $U_0 - U_M$ is a decreasing function of z, (34) implies that

$$U_M'(z) > U_0'(z) - \inf_{z \leqq z_0} U_0'(z) \geqq 0.$$

Therefore, z_M being defined by (28),

$$z_M > z_0. \tag{36}$$

Now (employing an argument already used for another purpose above)

$$\frac{\partial}{\partial z}\left[\int_0^\infty a_0(y) f(y, z) dy - g\left(\int_0^\infty y f(y, z) dy\right)\right] = -\int_0^\infty \{a_0'(y) - g'\} F_z dy < 0. \tag{37}$$

It follows from (37) and the definition of a_M that

$$\int_0^\infty a_M(y, z_M) dy - g\left(\int_0^\infty y f(y, z_M) dy\right)$$
$$\leqq \int_0^\infty a_0(y) f(y, z_M) dy - g\left(\int_0^\infty y f(y, z_M) dy\right)$$
$$< \int_0^\infty a_0(y) f(y, z_0) dy - g\left(\int_0^\infty y f(y, z_0) dy\right)$$
$$\leqq 0,$$

since the full-control optimum is feasible. Thus (27) is satisfied.

Finally, we have to show that $U_M(z_M) \to U_0(z_0)$ as $M \to \infty$. Using (35), we have

$$U_M(z_M) > U_M(z_0) = U_0(z_0) - \frac{A}{K}. \tag{38}$$

Also, since (as we have just seen) z_M is a feasible level of care when the income function is a_0,

$$U_M(z_M) < U_0(z_M) \leqq U_0(z_0). \tag{39}$$

Combining (38) and (39), and letting $K \to \infty$, we obtain, as desired,

$$U_M(z_M) \to U_0(z_0). \tag{40}$$

Thus we have proved (22). Since it was earlier established that the full-control optimum is not attainable, it now follows that the supremum is not attainable; and the proof is complete. ||

Because of the slightly awkward shape of the proof, adopted to ensure that z_M does indeed maximize expected utility, the crucial steps may be hidden. In effect, assumptions (i) and (ii) make it possible to introduce strong *marginal* work incentives with an arbitrarily small effect on total utility. Assumption (ii), leading to (34), implies that the marginal effect of penalties for high losses is disproportionately greater than the effect on expected utility. Assumption (i) (allowing (35) to be solved) then ensures that the marginal effect can be as large as desired. Assumption (iv) simply identifies the end of the loss-scale at which penalties should apply. If the opposite assumption, $u_{xy} \leqq 0$, holds, uncontrolled individuals take too much care rather than too little. If $\lim_{y \to 0} (f_z/f) = \infty$, large penalties for very small losses can then take the place of large penalties for very large losses, as in the proof. Assumption (iii) is technical; but severe violations would probably upset the result.

The theorem itself, though perhaps surprising, is not, as stated, disturbing. It is the argument leading to the theorem that is disturbing; the argument, in effect, that it is always possible to improve insurance arrangements yet further by increasing penalties for those with high losses (and correspondingly increasing the loss level above which the penalties apply). As K and M are increased without limit, the value of δ implied by (35) tends to zero. This is not apparent from the proof of the theorem. But if the steps leading to (35) are retraced, it will be seen that δ is so chosen that

$$\lim_{M \to \infty} \int_M^{\infty} \{u(a_0, y, z_0) - u(\delta a_0, y, z_0)\} f_z(y, z_0) dy \leqq A. \tag{41}$$

This is possible only if $\delta \to 0$. In fact, as the theorem states, any feasible income function that is positive for all y yields expected utility strictly less than the full-control optimum. In other words, *penalties of unlimited severity are essentially optimal.* It should be emphasized that this is true only under certain assumptions, and false under others. In particular unbounded utility is essential. Indeed, one can say that Theorem 1 is a form of the St. Petersburg paradox, which, as is well known, occurs only with unbounded utility. (Arrow has proposed that utility must be bounded for this reason, (1970*a*), Chapter 2; but this seems too easy a way of removing difficulties, and an unnecessary way of explaining why St. Petersburg contracts do not exist.) The cases where it is true are not intrinsically unreasonable. On the contrary, they seem to be readily applicable to many real situations. In later sections of the paper we shall discuss how robust the result is to certain important extensions of the model, and how seriously it should be taken.

In Section 5, the other class of cases, for which an optimum exists and finite losses are implied by unobservable behaviour, is discussed. First-order conditions are obtained, their validity established, and their interpretation explained. We also show how the two cases shade into one another, and how the striking feature of the first case, which can perhaps be called the penalty case, is not entirely absent in the second case. But before that, a new difficulty must be addressed.

4. NECESSARY CONDITIONS FOR AN OPTIMUM

The main technique for solving constrained maximization problems is Lagrange's method of undetermined multipliers. It was extended to problems with constraining inequalities by Kuhn and Tucker, and can be applied to functional problems in the Calculus of Variations as well as to problems in finite-dimensional spaces. Our problem, as stated in (13), is not of a type that has been previously treated. This section is devoted to a brief account of such a treatment, which emphasizes its peculiar difficulties.

Problem (13) is a special case of the following general situation:

$$\left.\begin{array}{l} a \text{ maximizes } U(a, z) \text{ subject to } z \text{ maximizing } V(a, z) \\ \qquad\qquad \text{and } H(a, z) = 0. \end{array}\right\} \tag{42}$$

In this paper, we are particularly interested in the case where a is a function; but first consider (42) with a and z finite-dimensional vectors, and U, V and H functions that are continuously differentiable at least twice. The set over which the maximization is taking place is possibly of rather awkward shape. This is illustrated by the following example.

Example 1. a and z are scalars.

$$\text{Find } a \text{ to maximize } -(a-2)-(z-1)^2,$$

$$\text{subject to: } z \text{ maximizes } V(a, z) = ae^{-(z+1)^2} + e^{-(z-1)^2}.$$

We begin by discovering the constraint set. The first-order condition for maximization with respect to z is

$$a(z+1)e^{-(z+1)^2} + (z-1)e^{-(z-1)^2} = 0,$$

i.e.

$$a = \frac{1-z}{1+z} e^{4z}. \tag{43}$$

This curve is shown in Figure 2: for a between 0·344 and 2·903 there are three values of z satisfying (43), and it still remains to discover which of them actually does the maximizing.

To settle this, we observe that

$$\begin{aligned} V(a, z) - V(a, -z) &= (a-1)(e^{-(z+1)^2} - e^{-(z-1)^2}) \\ &= -(a-1)(e^{4z} - 1)e^{-(z+1)^2}, \end{aligned}$$

so that for fixed $a > 1$, V is less for positive z than for negative; while if $a < 1$, V is greater for positive z than for negative. Therefore the maximum of V occurs for positive z when $a < 1$, for negative z when $a > 1$. In either case (as is readily verified) this identifies the desired solution of (43) uniquely. The points of the locus (43) for which z maximizes V are shown in the diagram as a heavy line; the remainder of the locus is light. When $a = 1$, V is maximized by $z = \pm 0{\cdot}957$.

It is clear, by sketching contours $(a-2)^2 + (z-1)^2 = \text{constant}$ in the diagram that the solution of the maximization problem is

$$a = 1, \qquad z = 0{\cdot}957.$$

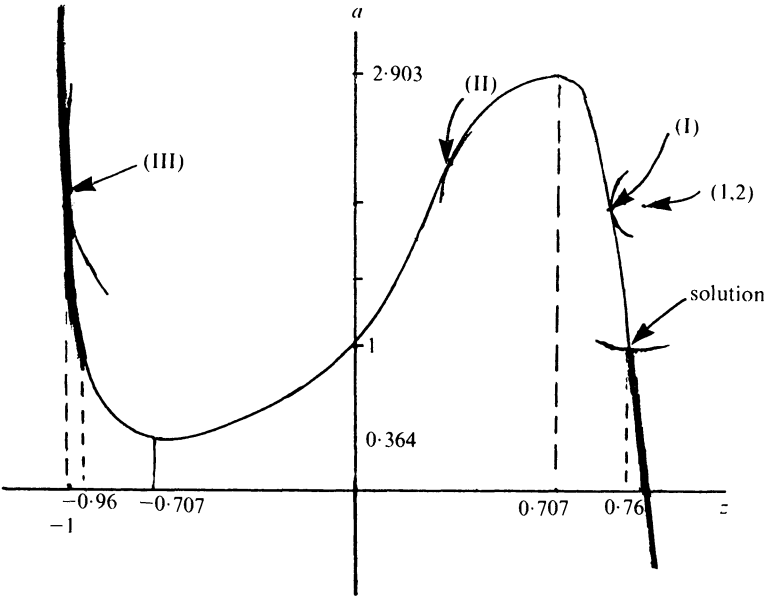

FIGURE 2

This solution is not obtained if one treats the problem as a conventional constrained maximization problem with the first-order condition (43) as contraint. The Lagrangian is then

$$-(a-2)^2-(z-1)^2+\lambda\left(a-\frac{1-z}{1+z}e^{4z}\right),$$

whose derivatives are zero when

$$2(a-2)=\lambda, \qquad 2(z-1)=\frac{4z^2-2}{(1+z)^2}e^{4z}\lambda, \qquad a=\frac{1-z}{1+z}e^{4z},$$

i.e. when

$$a=\frac{1-z}{1+z}e^4, \qquad 2a(2-a)=\frac{(1-z^2)^2}{(1+z)(2z^2-1)}.$$

There are three solutions:

(I) $a=1{\cdot}99$, $z=0{\cdot}895$;

(II) $a=2{\cdot}19$; $z=0{\cdot}420$;

(III) $a=1{\cdot}98$, $z=-0{\cdot}980$.

The first clearly gives the largest value for the maximand, $-(a-2)^2-(z-1)^2$, but our previous analysis shows that z does not maximize $V(a,z)$ for this value of a. As a matter of fact, z is a *local* maximum, but not a *global* maximum. The second solution is ineligible on all possible grounds: z is a local minimum of $V(a,z)$. The third solution, on the other hand, has the property that z is a global maximum of $V(a,z)$, so that it does satisfy the constraint of the original problem. But it is not the solution of that problem, and indeed gives a much lower value of the maximand than is actually possible.

This example shows that it is not legitimate to attempt to solve the problem by substituting first-order conditions for the maximization constraint. Furthermore, and this deserves emphasis, the example, though complicated, is in no sense special. Any moderate variation of the functions involved yields a problem with the same properties.

This inconvenient state of affairs arises because the set of (a, z) for which z maximizes a differentiable (even analytic) function $V(a, z)$ is not in general an open smooth manifold (as the set $V_z = 0$ is in general). This set, which we shall call the V-maximizing set, can consist of a number of disjoint closed components; and in that case the solution of the problem (42) is quite likely to lie on the boundary of one of these components. That is what happens in Example 1. Indeed difficulties arise whenever the V-maximizing set is a proper subset of the manifold defined by $V_z = 0$.

In order to be more precise, let us say that (a_0, z_0) is, by definition, *V-critical* if

$$\left.\begin{array}{l} z_0 \text{ maximizes } V(a_0, z), \text{ but there is no neighbourhood } N \\ \text{of } (a_0, z_0) \text{ such that } (a', z') \in N,\ V_z(a', z') = 0 \text{ imply} \\ \text{that } z' \text{ maximizes } V(a', z). \end{array}\right\} \tag{44}$$

(a_0, z_0) is certainly V-critical if the z_0 is an isolated but not unique maximum of $V(a_0, z)$. This is the usual case, as in Example 1, though there are other exceptional possibilities. In most problems that arise, one would find the V-critical points by looking for values of a at which the maximum is not unique.

The considerations mentioned above lead us to formulate an analogue of the Lagrange-multiplier method for the present problem:

Theorem 2. *Let U, V and H be twice continuously differentiable functions. If a^*, z^* maximize U subject to the constraints that z^* maximizes V and $H = 0$, and the matrix*

$$\begin{pmatrix} V_{zz}, & V_{za} \\ H_z, & H_a \end{pmatrix}$$

is of full rank at (a^*, z^*); then either (a^*, z^*) is V-critical, or there exists a vector λ and a scalar μ such that the derivatives of

$$L = U + V_z \cdot \lambda + \mu H,$$

with respect to a and z, vanish at (a^*, z^*).

Proof. The argument is routine.

Let da, dz be vectors such that

$$\left.\begin{array}{l} V_{zz} \cdot dz + V_{za} \cdot da = 0, \\ H_z \cdot dz + H_a \cdot da = 0, \end{array}\right\} \quad \text{at } (a^*, z^*). \tag{45}$$

By the full-rank condition, and the implicit function theorem, (45) implies that there exist continuously differentiable functions $z(s)$ and $a(s)$, of a scalar s varying between -1 and

1, such that

$$\left.\begin{aligned} &V_z(a(s), z(s)) = 0, \qquad H(a(s), z(s)) = 0, \quad (|s| \leqq 1), \\ &a'(0) = da, \qquad\qquad z'(0) = dz, \\ &a(0) = a^*, \qquad\qquad z(0) = z^*, \end{aligned}\right\} \tag{46}$$

if (a^*, z^*) is not V-critical, $V_z(a(s), z(s)) = 0$ implies that, for all sufficiently small s, $z(s)$ maximizes $V(a(s), z)$. Therefore, when s is small enough, $a(s), z(s)$ satisfy the constraints of the maximization problem, and consequently

$$U(a(s), z(s)) \leqq U(a^*, z^*) = U(a(0), z(0)).$$

Hence at $s = 0$ the derivative of $U(a(s), z(s))$ with respect to s vanishes

$$0 = U_a \cdot a'(0) + U_z \cdot z'(0) = U_a \cdot da + U_z \cdot dz. \tag{47}$$

Since (47) is implied whenever (da, dz) satisfy (45), there exist a vector λ and a scalar μ such that

$$\left.\begin{aligned} U_a &= -\lambda \cdot V_{za} - \mu H_a, \\ U_z &= -\lambda \cdot V_{zz} - \mu H_z. \end{aligned}\right\} \tag{48}$$

This is the stated result. ||

The full-rank condition is, of course, a generic one, which will fail to be satisfied only in exceptional cases. But, as we saw in Example 1, V-criticality is not exceptional. We must therefore devote some attention to conditions that are satisfied by a V-critical optimum. This discussion will be somewhat heuristic, to avoid tedious details.

Normally, a V-critical point (a', z') is such that there are two distinct maxima, z' and z'', of $V(a, \cdot)$, there being no others. Furthermore, a normally belongs to a critical boundary in a-space, a boundary consisting of a which are critical (points on either side of the boundary having a unique maximizing z corresponding to them). This boundary is (again, normally) $(m-1)$-dimensional when a is m-dimensional, and locally describable by a single direction vector, normal to the boundary. Since, along this boundary,

$$V(a, z') = V(a, z''),$$

and also

$$V_z(a, z') = V_z(a, z'') = 0,$$

we have, by the envelope theorem

$$\{V_a(a', z') - V_a(a', z'')\} \cdot da = 0,$$

for any direction da in the boundary. Therefore the normal p to the boundary can be taken to be

$$p = V_a(a', z') - V_a(a', z'').$$

When $p \cdot da > 0$, V increases by more in the part of $V_z = 0$ near (a', z') than in the part near (a', z''). Therefore when a is close to a', and $p \cdot a > p \cdot a'$, the maximizing z is close to a'.

We can therefore modify the argument used in proving Theorem 2, and claim that, when (a^*, z^*) is V-critical, and p is defined as

$$p = V_a(a^*, z^*) - V_a(a^*, z''), \tag{49}$$

(where z'' is the other maximizing value of z for a^*),

$$V_{zz} \cdot dz + V_{za} \cdot da = 0, \qquad H_z dz + H_a \cdot da = 0, \qquad p \cdot da \geqq 0, \tag{50}$$

together imply that

$$U_z \cdot dz + U_a \cdot da \leqq 0, \tag{51}$$

because (50) implies the existence of a path satisfying the constraints. It then follows by the Minkowski–Farkas lemma that there exist a vector λ, a scalar μ, and a non-negative scalar ν, such that

$$\left.\begin{aligned} U_z &= -\lambda \cdot V_{zz} - \mu H_z, \\ U_a &= -\lambda \cdot V_{za} - \mu H_a - \nu p. \end{aligned}\right\} \tag{52}$$

(52) can be stated in Lagrangian form, by saying that the derivatives of

$$L = U + V_z \cdot \lambda + H\mu + \{V(a, z) - V(a, z'')\}\nu, \tag{53}$$

with respect to a and z are zero at (a^*, z^*) (z'' being a distinct maximum of $V(a^*, \cdot)$). Furthermore, $\nu = 0$ when a^*, z^* is not critical. It turns out, therefore, that our problem can be solved in the Kuhn–Tucker manner if we replace it by:

$$\left.\begin{aligned} &a^*, z^* \textit{ maximizes } U(a, z), \textit{ subject to: } V_z(a, z) = 0, \\ &V(a, z) \geqq V(a, z'') \textit{ for all } z'' \textit{ such that } V_z(a, z'') = 0, \\ &H(a, z) = 0. \end{aligned}\right\} \tag{54}$$

Seen from this point of view, conditions (52) (in conjunction with (50)) seem quite intuitive.

These conditions would hardly have been of much use to us in solving Example 1, because there both z and a were 1-dimensional, so that the V-critical set consisted of isolated points. Once these were found, there was little more work to do. In other problems of higher dimension, one would have to resort to conditions (52) to obtain the solution. But it must be emphasized that identification of the V-critical set is an essential preliminary; which, in general, looks far from easy. In our next example, we have again, by a trick, reduced matters to the convenience of two dimensions.

Since the work of this section has been couched in very general terms, it is not clear as yet whether it has much or little relevance to the moral hazard problem. Unfortunately, it is highly relevant: there is no great difficulty is constructing examples of moral hazard problems in which the solution is a critical point, and the simple Lagrangian technique is therefore invalid.

Example 2. Suppose only two output levels are possible, called 1 and 0, and that the utility of income and effort in these two cases is

$$u_1(a, z) = 3(1 - e^{-4a}) - z + \tfrac{1}{10} \arctan(10z),$$

$$u_0(a, z) = 5a^{1/5} - z + \tfrac{1}{10} \arctan(10z).$$

The probabilities of outputs 1 and 2 are

$$p(z) = (1+z)e^{-z},$$

and

$$1-p(z).$$

We wish, if possible, to choose x_1 and x_2 so as to maximize

$$U = pu_1 + (1-p)u_0,$$

subject to: z maximizes $pu_1 + (1-p)u_2$

$$px_1 + (1-p)x_2 = 1.$$

Interpretation. p is the probability of being involved in an accident, which varies between 0 and 1 according to the amount of care taken. The effect of being involved in an accident is to reduce utility. It so happens that the marginal utility of income is increased between $a = 0{\cdot}06$ and $a = 0{\cdot}47$, but this special (though not unreasonable) feature of the example plays no role in what follows. The disutility of care is not affected by the accident: this represents the case of precautions taken before the accident may occur. The problem is, then, a perfectly plausible one.

Solution. First we find the maximizing set for the agent. His first-order condition is

$$ze^{-z}\{5a_0^{1/5} - 3(1-e^{-4a_1})\} = \frac{100z^2}{1+100z^2}.$$

Writing $w = u_0 - u_1 = 5a_0^{1/5} - 3(1-e^{-4a_1})$, we have

$$w = \frac{100ze^z}{1+100z^2}. \tag{55}$$

This curve is shown in Figure 3. By direct computation, it is found that z is the utility-maximizing choice corresponding to w when (z, w) is on the heavy part of the curve. The critical points are given by $w = 3{\cdot}055$ and $z = 0{\cdot}034,\ 1{\cdot}575$.

By using the last constraint, we can express total utility as a function of w and z

$$U = 5a_0^{1/5} - wp(z) - z + \tfrac{1}{10}\arctan(10z), \tag{56}$$

with a_0 given as a function of w and z by

$$-\tfrac{1}{4}\log\{1 - \tfrac{1}{3}(5a_0^{1/5} - w)\} = a_1, \qquad a_1 p + a_0(1-p) = 1. \tag{57}$$

From (56) and (57), we find the derivatives of U with respect to w and z when (w, z) lies on the maximizing locus

$$\frac{\partial U}{\partial z} = a_0^{-4/5}\frac{\partial a_0}{\partial z} = \frac{a_0^{-4/5}(a_1 - a_0)ze^{-z}}{Bp + 1 - p}, \tag{58}$$

where $B = \frac{1}{12}a_0^{-4/5}\{1 - \frac{1}{3}(5a_0^{1/5} - w)\}^{-1} = \frac{1}{12}a_0^{-4/5}e^{4a_1}$, and

$$\frac{\partial U}{\partial w} = a_0^{-4/5}\frac{\partial a_0}{\partial w} - p = \frac{p(1-p)(B-1)}{Bp + (1-p)}. \tag{59}$$

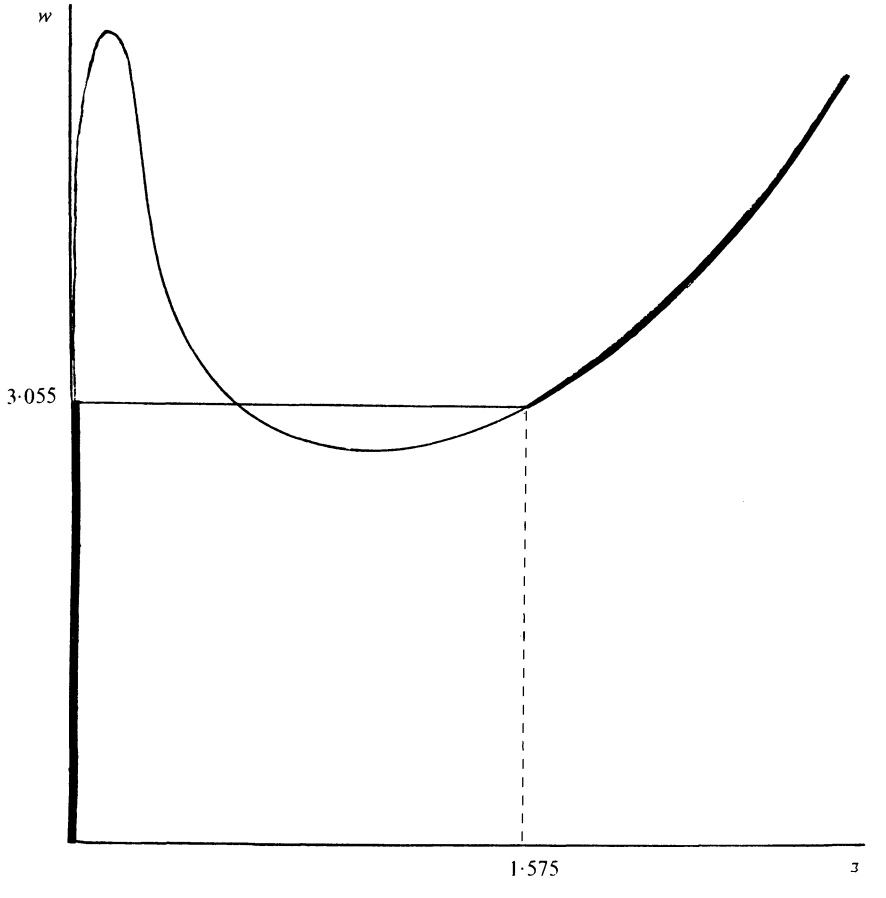

FIGURE 3

Given w and z, a_0 and a_1 are determined by $a_1p + a_0(1-p) = 1$ and the definition of w. Since $a_1 \geqq 0$ and $w \leqq u_0(a_0)$, this imposes an upper bound on w and z, which need not concern us. So long as $w > u_0(1) - u_1(1) = 2{\cdot}055$, $a_0 > a_1$; for if $a_0 \leqq a_1$, $a_0 \leqq 1 \leqq a_1$ (since $1 = a_1p + a_0(1-p)$) and $w = u_0(a_0) - u_1(a_1) \leqq u_0(1) - u_1(1)$. When $w \leqq 2{\cdot}055$, $a_0 \leqq 1 \leqq a_1$. $z = 0{\cdot}021$ when $w = 2{\cdot}055$. Therefore, on the maximization set, when $z \leqq 0{\cdot}021$, U_z and U_w are both positive; since there $a_0 < a_1$, and $B = \frac{1}{12}a_0^{-4/5}e^{4a_1} > \frac{1}{12}e^4 > 1$.

Consider next the part of the maximization set with $0{\cdot}021 < z \leqq 0{\cdot}034$. Here $U_z < 0$ and $U_w < 0$: to determine how U varies as we move along the curve, we have to consider

$$U' = U_w \frac{dw}{dz} + U_z.$$

Throughout this part of the curve, dw/dz is never less than its value at $z = 0{\cdot}034$, which is calculated to be 75·3. a_1 here decreases as z and w increase along the curve, and equals 0·999 at $z = 0{\cdot}034$; and a_0 increases from 1 to 2·488 at $z = 0{\cdot}034$. Thus $a_1 - a_0 > -1{\cdot}489$. Also $ze^{-z} < 0{\cdot}033$, its value at $z = 0{\cdot}034$; and $p(1-p) > 5{\cdot}65 \times 10^{-4}$, its value at 0·034.

Therefore

$$U' > (Bp+1-p)^{-1}\{5{\cdot}65\times 10^{-4}(\tfrac{1}{12}a_0^{-4/5}e^{4a_1}-1)\times 75{\cdot}3 - 1{\cdot}489\times 0{\cdot}033a_0^{-4/5}\}$$
$$= (Bp+1-p)^{-1}a_0^{-4/5}(3{\cdot}54\times 10^{-3}e^{4a_1} - 0{\cdot}043a_0^{-4/5} - 0{\cdot}049)$$
$$> (Bp+1-p)^{-1}a_0^{-4/5}(3{\cdot}54\times 10^{-3}\times 54{\cdot}38 - 0{\cdot}043\times 2{\cdot}07 - 0{\cdot}049),$$

and the numerical expression comes to 0·054, which is positive. As far as the left hand branch of the curve is concerned, then, we find that the critical point $w = 3{\cdot}055$, $z = 0{\cdot}034$ gives the largest value of U.

In the case of the right-hand branch ($z \geqq 1{\cdot}575$), $a_0 > a_1$, as we have already seen, a_1 decreases along the branch (because a decreasing function of both w and z), and is found by calculation to be 0·433 at the critical point. Therefore

$$B = \tfrac{1}{12}a_0^{-4/5}e^{4a_1} < \tfrac{1}{12}e^{1{\cdot}732} = 0{\cdot}471 < 1.$$

This means that U_z and U_w are both negative on the right hand branch, so that U decreases on its as z increases. It follows from these considerations that the solution of the problem is at a critical point, since it cannot be at any other point of the V-maximizing locus. Of the two critical points, the left-hand one is the better, since a_0 is found by direct calculations to be larger there (2·488 as compared to 1·646). The solution to the problem is therefore

$$a_0 = 2{\cdot}488, \qquad a_1 = 0{\cdot}999, \qquad z = 0{\cdot}034.$$

The solution is a surprising one, for it involves a very high risk of an accident, with the agent insuring against the risk of not having an accident and therefore having an enhanced marginal utility of income. It is not without interest that such a solution is possible in a model that is not entirely unreasonable. When in Section 6, we discuss this general class of models, it will be seen that the condition implying reverse insurance (against the risk of not suffering loss) is not the one that we would, at first thought, guess.

The main point of the example is that, once again, the substitution of first-order conditions for the maximization constraint is invalid. The more general treatment in Section 6 will make it clear that the difficulty is not exceptional. The solution must also have shown how extensive are the calculations apparently required for quite a simple problem. Matters would have been even more trying if it had not been possible to reduce the analysis to a two-dimensional diagram. The moral hazard problem initially posed in this paper is an infinite-dimensional one, with a a function, not a finite-dimensional vector. This is discouraging for general analysis of the problem. But there are good reasons for considering the choice of a general payment function. The result of Section 1 gave one such reason. The other reason is explained in Section 5.

Acknowledgements. Earlier versions of some of the methods and results of this paper were presented at seminars at Oxford, L.S.E., Cambridge, Edinburgh and Warwick. I am grateful for numerous comments at these seminars. Financial support by the National Science Foundation is gratefully acknowledged.

REFERENCES

ARROW, K. J. (1970*a*), "Alternative Approaches to the Theory of Choice in Risk-Taking Situations", in *Essays in the Theory of Risk Bearing* (Amsterdam: North Holland).

ARROW, K. J. (1970*b*), "The Economics of Moral Hazard: Further Comment", in *Essays in the Theory of Risk Bearing* (Amsterdam: North Holland).

CHEUNG, S. N. (1969) *The Theory of Share Tenancy* (Chicago: Chicago University Press).

MIRRLEES, J. A. (1972), "Population Policy and the Taxation of Family Size", *Journal of Public Economics*, **1**, 169–198.

MIRRLEES, J. A. (1974), "Notes on Welfare Economics, Information and Uncertainty", in M. Balch, D. McFadden and S. Wu (eds.) *Essays in Equilibrium Behavior under Uncertainty* (Amsterdam: North Holland).

MIRRLEES, J. A. (1976), "The Optimal Structure of Incentives and Authority within an Organisation", *Bell Journal of Economics*, **7**, 105–131.

PAULY, M. V. (1968), "The Economics of Moral Hazard", *American Economic Review*, **58**, 531–537.

ROSS, S. (1973*a*), "The Economic Theory of Agency: the Principal's Problem", *American Economic Review*, **63**, 134–139.

ROSS, S. (1973*b*), "On the Economic Theory of Agency", Proceedings of the NBER Conference on Decision Making and Uncertainty.

RADNER, R. (1968), "Competitive Equilibrium under Uncertainty", *Econometrica*, **36**, 31–58.

SPENCE, M. and ZECKHAUSER, R. (1971), "Insurance, Information, and Individual Action", *American Economic Review*, **61**, 380–387.

STIGLITZ, J. E. (1974), "Incentives and Risk Sharing in Share Cropping", *Review of Economics Studies*, **41**, 219–256.

STIGLITZ, J. E. (1975), "Incentives, Risk and Information: Notes Towards a Theory of Hierarchy", *Bell Journal of Economics*, **6**, 552–579.

ZECKHAUSER, R. (1970), "Medical Insurance: A Case Study of the Tradeoff Between Risk Spreading and Appropriate Incentives", *Journal of Economic Theory*, **2**, 10–26.