1985

Rules And Institutions

Philip Nicholas Rowe

CANADIAN THESES ON MICROFICHE

I.S.B.N.

THESES CANADIENNES SUR MICROFICHE

National Library of Canada
Collections Development Branch

Canadian Theses on
Microfiche Service

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Direction du développement des collections

Service des thèses canadiennes
sur microfiche

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C 30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C 30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

RULES AND INSTITUTIONS

by
P. Nicholas Rowe

Department of Economics

Submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario
February 1985

ABSTRACT

This thesis contains four related but self-contained essays, plus a short introductory chapter. The central theme is the rationality of agents' following rules, as opposed to discretionary action, in contexts where optimal plans under rational expectations are time-inconsistent. The first essay examines the rationality of keeping promises and of trusting that promises will be kept. The second essay argues that social institutions are identical to agents' following, and being believed to follow, rules of action, and uses this perspective to analyse the existence of property rights. The third essay uses the assumption of costly monitoring of worker malfeasance and hence imperfect trust to explain non-compensating wage-differentials between identical workers and equilibrium unemployment. The fourth essay presents a theory of strikes as the consequence of the rules of action followed by firm and worker to enforce a contingent wage contract under imperfectly symmetric information.

ACKNOWLEDGEMENT

My major debt is to David Laidler, chairman of my advisory committee, for encouraging me in my pursuit of the slightly unorthodox lines of enquiry which eventually lead to this thesis. I am also indebted to the other members of my committee - Joel Fried, Glenn MacDonald, Ron Wintrobe, and (temporarily) Chris Robinson - for their advice and for forcing their (sometimes unwilling) student to express his ideas with greater clarity and rigour. Of the many other people to whom I am grateful for discussion and comments, I thank Tim Lane in particular for the value and extent of his help. My wife Rosslyn Emmerson, provided not only moral but also intellectual support in completing this thesis. I thank the University of Western Ontario and the Ontario taxpayer for providing financial assistance to a foreign graduate student, Carleton University for paying typing expenses, and Tara McCreery for efficiently and patiently typing the final manuscript.

TABLE OF CONTENIS

Chapter One

Introduction

This thesis consists of four essays:

1. "The Rationality of Promises"
2. "An Economic Explanation of the Nature and Existence of Social Institutions"
3. "Trust, Wages and Employment"
4. "An Equilibrium Theory of Strikes"

Each essay can be read independently, but the full implications of each essay can better be understood if they are read together. This introduction briefly describes the contents of each essay, and explains how they are related.

The second essay is the key to the thesis. The first essay provides a foundation for the concepts and arguments used in the second essay. The third and fourth essays explore extensions and applications of the basic approach developed in the second essay.

The second essay seeks to explain what social institutions are, and why they exist. The term "social institution" is meant to cover such things as law, language, governments, nations, systems of etiquette, bidding procedures, monetary exchange, contracts, clubs, corporations, etc., but for simplicity the essay discusses one particular type of social institution, of obvious relevance for economics - a system of property rights.

It might be argued that an economic theory of the nature and existence of property rights already exists. Demsetz (1967, p. 350), for example, has argued that "...property rights develop to internalize externalities when the gains of internalization become larger than the cost of internalization." This theory is a particular instance of a general approach which argues that the social institutions which exist do

so because they lead to a more efficient allocation of resources than any alternative. Given the validity of this approach, it would suffice merely to spell out the benefits and costs of alternative institutional arrangements, showing that the net benefits of the existing institution exceed that of its potential competitors, to claim to have explained why that institution exists.

This general approach is known as Functionalism. One explains the existence of an institution by explaining its function. Economists who follow this approach are merely functionalists who assume that efficiency is the relevant functional criterion.

Any functionalist theory is incomplete in that by explaining the consequences of the existence of an institution one does not ipso facto explain the causes of the existence of that institution.[1] Any functionalist explanation must be supplemented by a theory which links those consequences with the cause. One candidate for such a link would refer to the intentions of the agents who created the institution. Realising that all would be better off under some system of property rights, agents collectively agree to create such a system. A second candidate for such a link would refer to some selection process. Less efficient institutions tend to lose members to more efficient competitors, so that only the most efficient survive.[2]

There are problems with both candidates for a theory to link consequences with causes. The intended-consequence explanation seems to presuppose a high degree of wisdom and foresight on the part of the founders of an institution. The survivalist explanation seems to ignore the possibility that social institutions might be natural monopolies; the existence of the current institutions may preclude the creation of a more efficient competitor. A system of private and of common property rights

cannot coexist on the same area of land. No one agent can leave to join a hypothetical competitive institution. Either all leave together, or else nobody does.

If these criticisms are accepted then the functionalist account loses much of its validity. The continued existence of current institutions may result as much from inertia due to ignorance of alternatives or to natural monopoly as it does from their efficiency. Although it is always possible to save the axiom that existing institutions are efficient by invoking "costs" of change, much of the axiom's simplicity and empirical content would thereby be lost.

These criticisms aside, however, there is a more fundamental inadequacy in the functionalist approach. The functionalist account views an institution as defining a set of constraints on the behaviour of individual members of that institution. Their behaviour subject to these constraints leads to an allocation of resources more efficient than would their behaviour subject to a different set of institutional constraints, or subject to no institutional constraints at all.[3] The fallacy lies in thinking of these constraints as existing independently of the actions and beliefs of the individual agents who act subject to those constraints. Institutional constraints are conceptually distinct from the bricks and mortar of the constraints of our physical environment. That all would gain if all acted within the institutional constraints does not entail that each would gain if he were to do so.

Who defines the constraints? Who enforces them? Why do they do so? These are the questions which must be addressed if we seek to explain the existence of social institutions. It is no answer to suggest that property rights are defined and enforced by the state. The state is itself a social institution, the existence of which cannot be presupposed if we

seek to reconcile the existence of social institutions with the methodological assumption of rational individual action and belief. Who defines the state? Who enforces the definition?

Social institutions do not exist independently of the actions and beliefs of the individual agents who comprise the membership of that institution. If no one believes in the existence of property rights, or makes any behavioural distinction between his and others' property, then property rights do not exist. But then what is it that an agent believes when he believes that something is not his property? What is it that he does when he violates another's property rights? Why does he hold these beliefs? Why does he make this behavioural distinction? The institutional constraints are not imposed exogenously on the game played by rational individual agents. The existence of the institution is instead a feature of the outcome of the game - some description of the beliefs held and actions taken by the players in the game's equilibrium. What sort of beliefs and actions are these that constitute the existence of a social institution? Is it rational for individual agents to hold those beliefs and perform those actions? These are the questions addressed in the second essay of the thesis. Only an answer to these questions can fully explain the nature and existence of social institutions.[4]

To answer these questions, the second essay posits an initial "state of nature" - a state of affairs in which all social institutions are absent. Building on previous analysis,[5] we explain why life in the state of nature is "nasty, brutish, and short." Mere revelation of the inefficiency of the state of nature compared to an allocation in which agents face exogenously imposed constraints does not suffice, however, to explain why social institutions exist. To explain why property rights exist we must show that the state of nature is not an equilibrium state -

that individual actions and beliefs therein are not chosen rationally. We argue that individuals in the state of nature are irrational in failing to follow rules of action.

In eschewing discretionary action in favour of following a rule, the individual sacrifices his freedom to act so as to maximise his utility given the (rational) expectations of others, in order to influence expectations of others via his choice of the rule of action to which he commits himself. The performance of an action specified by an optimal rule is generally irrational except insofar as its performance is required to maintain other's expectations that that rule will be followed in future. The equilibrium wherein individuals rationally follow such rules of action, and rationally believe that others do likewise, it is argued, can justifiably be said to be an equilibrium in which social institutions exist. Social institutions are thus theoretical constructs, shared by social theorist and social agent alike, which permit "as if" explanations of agents' action when agents follow, and are believed to follow, such rules of action. The mistake of functionalism is its granting these theoretical constructs a concrete and independent existence. Social institutions are not features of the environment; they are theories about what agents believe and how they behave.

The approach to the study of social institutions adopted in the second essay rests on a distinction between discretionary action and following a rule. This distinction is developed in the first essay under the guise of the following question: "Why do agents ever keep their promises?"

An agent who is believed to keep his promises can influence another's expectation concerning his future actions by promising to perform those future actions. By thus influencing the other's expectations,

he can influence the other's current action. When tomorrow arrives, however, the aims an agent sought to achieve by promising have already been attained. Why should he keep his promise? Why should others rationally expect him to do so? The problem is that of the time-inconsistency of optimal plans under rational expectations.[6] The future action to which I would today optimally reveal myself committed does not generally coincide with that specified by my future optimal plan. Could it ever be rational to follow the rule specified by one's past optimal plan? Could it ever be rational to conform to constraints which everyone knows to be purely imaginary? Could it ever be rational to believe that another would do so?

Assuming that agents do not have a <u>taste</u> for keeping promises, the only reason an agent might have of keeping a promise is if his doing so or failing to do so in this instance might influence others' expectations concerning whether he will keep future promises. If it could be known a priori whether or not an agent will keep his future promises then he would have no incentive to keep his current promise. This demonstrates that there exists one equilibrium in which it is known that agents will never, under any circumstances, attach any importance to their having promised. Knowing that everybody knows they will never keep their promises, they never do, and the social institution of promising does not exist. However, it is always an option for an agent with free will to destroy this equilibrium by keeping a promise and thereby destroying the possibility of having a priori knowledge he forces them to look to his past compliance with promises as a guide to his future compliance, just as he intends. Realising this intention, others realise that he now has an incentive to keep his promises in order to maintain their belief that he will do so in future, which again, is just what he intends them to

realise. It can, therefore, be rational to follow the rule of promise-keeping, and hence can be rational to believe that another might do so, though this belief can never be a priori knowledge.

In following a rule an agent forgoes his freedom to maximise his utility-subject to the natural informational environment, in order to be free to create an artificial informational environment via his revealed commitment to a rule. It is this creation of artificial informational environments which lies at the heart of all social institutions - of which the institution of promising is the most basic.

Exchange is the exchange of property rights, and therefore presupposes a system of property rights. A system of property rights, like any social institution, is maintained in existence only by agents' following rules of action. An agent will rationally follow a rule only if the value of maintaining his reputation for doing so exceeds the benefits he could obtain by departing from the rule. The first and second essays assume that the values of agents' reputations can be sufficient never to impose a binding constraint on the form of rule an agent might credibly follow. This assumption cannot hold generally. If an agent gains little from being believed to keep his promises, if he can gain a lot from violating a promise, if the probability of a violation's being detected is low, and if the future benefits from having a reputation are discounted at a high rate, then the promises which the agent would like to make may not be credible. The third essay in the thesis: "Trust, Wages, and Employment," illustrates the implications of limited trust on the prices and quantities of goods exchanged.

We consider a labour market with a fixed number of identical, anonymous workers. An employed worker has opportunities for "cheating" his employer - undertaking some activity such that the costs of that activity

to the employer exceed the benefits to the worker. Since "cheating" imposes net costs, any worker would like ex ante to commit himself to refraining from "cheating" in return for the higher wages his employer would be willing to pay. Could such a commitment be credible? We assume that monitoring is costly and hence imperfect, and that the only penalty the employer can inflict on a cheating worker is to fire him. The worker will not post a bond because he cannot trust his employer not to abscond with the bond.

If wages are set at market-clearing levels, then the worker's commitment is never credible. A worker caught "cheating" his employer and fired can immediately get another job at the same wage, and so loses nothing. In order to deter cheating, it may under specified circumstances pay an employer to raise wages above the supply price of labour to his firm, since workers would then lose something if their "cheating" is detected and they are fired. In the absence of those specified circumstances it may pay another employer to permit the inefficient activity "cheating," since his costs of monitoring, and of paying wages sufficiently above the supply price of labour to deter cheating, may exceed the net costs of "cheating." If some firms deter "cheating" by paying a wage above their supply price of labour, while other firms do not, the result is a wage differential which does not preserve equality of net advantage between jobs. If all firms deter "cheating" then, since all wages are above the supply price of labour, the result is an underemployment equilibrium. No firm will cut wages because if it were to do so it would fail to deter its workers from "cheating."

The importance of the third essay is that it shows that the fact that social institutions are not exogenously imposed concrete constraints on agents' behaviour can have implications for the behaviour that takes

place within any feasible institution - implications for such standard economic phenomena as prices and quantities traded.

The fourth essay, "An Equilibrium Theory of Strikes" is again an application and extension of the apprach to the analysis of social institutions developed in the second essay. For concreteness, the essay is introduced as presenting a theory of strikes, but the approach is intended as being applicable to all forms of conflict over the distribution of rights.

Any equilibrium theory of conflict - which seeks to reconcile the existence of costly conflicts with the assumption of rational action and belief - necessarily faces the following puzzle; since conflict is costly, there must exist some agreement which could make both sides better off than if they continue the conflict. Why then should rational agents sometimes fail to reach such an agreement immediately? One answer to this puzzle might be to renounce the attempt to form an equilibrium theory of conflict - to concede that conflicts result from "mistakes."[7] In renouncing equilibrium theory, however, we may renounce the possibility of any further understanding of conflicts in terms of agents' having reasons for acting as they do. If "mistakes" are possible, then is not any sort of "mistake" possible? Can we put any content into a theory of why and when agents make "mistakes?" My answer to this puzzle is not to reject the equilibrium approach as a valid approach to the study of conflict, but to apply a different sort of equilibrium approach. The equilibrium theorist of conflict must think not in terms of rational actions, but in terms of rational rules of action.

Any social institution, I have argued, such as a system of property rights or an implicit contract between firms and worker, is constituted and maintained by the rules of actions of the agents who comprise that

institution. An implicit contract owes its existence to, or rather, is the very same thing as, the belief that agents follow rules of action requiring them to punish and hence deter violations of the contract. During the normal course of affairs, the existence and definition of social institutions is so much a "taken-for-granted," both for economic agents and for economists, that they are treated as if concrete and independently existing entities, while the rules of action which constitute those institutions remain hidden from view. The implicit contract is enforced by the belief that violations will be punished by a withdrawal from trade - a strike - but normally the threat of strikes suffices to deter violations and ensures that actual strikes are not observed. Only in "abnormal" circumstances, such as when the definition of the social institution is in dispute, is it no longer possible to consider social instiutions as if concrete. Only then are revealed the rules of action which reconstitute and preserve the normal state of affairs.

Consider an implicit contract between firm and worker which sets the wage contingent on certain variables. If the firm and worker have imperfect and independent information on those variables they may differently estimate the contractually specified wage with the results that each perceives the other to be in violation of the contract. Since they are unable to distinguish a priori between truthful and untruthful announcements concerning the contractual wage, each must punish all perceived violations of the contract if the contract is to be enforced. The puzzle of the apparent irrationality of behaviour in a strike is now resolved, for the very definition of following a rule of action implies that an agent can gain from violating his rule if only he could do so without destroying his reputation for following that rule in future. Both sides

could gain by reaching an agreement to end the strike, but if it were anticiapted that they would act thus then both would lose, for the contingent contract could then no longer be enforced.

An explicit formal model is presented to illustrate the theory of conflict which demonstrates the possiblity of the existence of strikes. The equilibrium theory is shown capable of yielding strict predictions about the behaviour of firm and worker during a strike.

The four essays that comprise this thesis thus present a coherent approach to the study of human action and belief as it pertains to the creating and maintaining of the social institutions which define the context within which economic activity takes place. It is both necessary and possible to reconcile the existence of social institutions with the economic methodology of rational individualism, but is possible only if we shift our attention from the rationality of actions to the rationality of rules of action.

Footnotes

1. A similar criticism of Functionalism in sociology can be found in any sociology textbook, e.g., Turner (1978, p. 105).

2. "The reason why such rules will tend to develop is that the groups which happen to have adopted rules conducive to a more effective order of actions will tend to prevail over the groups with a less effective order." Hayek (1973, p. 99).

3. This seems to characterise the approach taken by e.g., Schotter (1981).

4. So far as I am able to ascertain, it is the asking of this sort of question, and the sort of approach which results from asking such questions, which characterises the influential work of Berger and Luckman (1966).

5. Hobbes' Leviathan obviously, but also Skogh and Stuart (1982).

6. See Kydland and Prescott (1977).

7. As in Hicks (1964).

References

Berger, P.L. and T. Luckman (1966), The Social Construction of Reality, New York.

Demsetz, H. (1967), "Towards a Theory of Property Rights", American Economic Review.

Hayek, F.A. (1973), Law, Legislation and Liberty Vol. I. University of Chicago Press; Chicago.

Hicks, J.R. (1964), The Theory of Wages, 2nd ed., London: MacMillan.

Hobbes, T. (1969), Leviathan. Bobs-Merill, 1958.

Kydland, F.E. and E.C. Prescott (1977), "Rules Rather than Discretion: The Inconsistency of Optimal Plans," Journal of Political Economy.

Schotter, A. (1981), The Economic Theory of Social Institutions, Cambridge University Press.

Skogh, G. and C. Stuart (1982), "A Contractarian Theory of Property Rights and Crime," Scandinavian Journal of Economics, 84.

Turner, J.H. (1978), The Structure of Sociological Theory, revised ed., Dorsey Press, Homewood, Illinois.

Chapter Two

The Rationality of Promises

I. Introduction

Alan wants to borrow money. He goes to Bill, who lends him $100 in return for a promise from Alan to repay the loan (with interest) next year. Next year Alan repays the loan as promised.

Nothing in this simple story sounds remarkable to the lay observer. It is the sort of thing that can happen all the time. It is more the exceptions that are remarkable and stand in need of explanation, when Bill fails to accept the offer to lend money, or when Alan fails to repay the loan. This attitude carries over to economists. The making of loans is unremarkable. They are simply exchanges, and are the result of the same sort of differences in preferences and opportunities that explain all exchanges. It is the exceptions which are puzzling, abnormal, and in need of special explanation. Why do credit-constraints and bankruptcies exist?

This attitude implies a complete inversion of the questions that should be asked. What is puzzling and in need of explanation, or ought to be seen as such from the economic approach to rational action and belief, is not that loans are sometimes refused, and sometimes not repayed, but that they are ever made and repayed at all.

Why is it rational for Alan to give Bill $110 in year two? If he does so he merely makes himself $110 worse off. Is it altruism which makes him do so? If it were so he would be as likely to give Bill $110 if he had found the sum lying in the gutter. Whatever reason could Bill have for giving Alan $100 in year one? Whatever reason could he have for believing that his giving Alan $100 in year one might be causally related to Alan's giving Bill $110 in year two? Surely Bill knows that Alan is a rational economic agent, for whom bygones are bygones. How can Bill's

giving Alan money in year one influence the rationality of Alan's giving Bill money in year two, except in the same way as would Alan's finding $100 in the gutter? The story is obviously preposterous. It involves an irrational action on Alan's part and hence an irrational belief on Bill's part.

This, I submit, is how an economic theorist ought to approach this story. I suggest that this is how an economic theorist would approach the story if he had never had the direct experience of such phenomena which enables him to take it for granted.

The statist is not puzzled. He agrees that the story would indeed be preposterous, and the economic analysis quite correct, were it not for the fact that the state intervenes either to force repayment of loans or to threaten penalties sufficient to make voluntary repayment the rational choice. Indeed the function of the state is to do just this sort of thing. State intervention makes the loan possible and makes both people better off. This is why the state exists.

There is, of course, an element of truth in the statist position. Sometimes loans will not be made without making use of third parties to ensure repayment. The introduction of third parties can thus make both parties better off, but this is not the end of the problem. Sometimes loans are made which cannot or will not be enforced by third parties. Even when loans are enforced by third parties, however, it is not legitimate simply to introduce third parties as a deus ex machina. The third party is not a mechanism, but must be analyzed as a rational agent with interests and beliefs of his own. Why is it rational for the third party to enforce the loan in year two? What profit can be made from the transfer of $110 from one agent to another? There is no potential pareto improvement to be found. Alan would pay as much to prevent the transfer

as Bill would to effect it. Supposing the third party to be "the state" does not help either, if "the state" is no more than a particular set of explicit and implicit contractual relations between cognitive, interested individuals. It is made of the same stuff that binds the loan agreement, and cannot therefore be presupposed in explaining the possibility of the latter. The statist explanation, at best, is incomplete. At worst, it begs the very questions it purports to answer.

The layman, like the statist, is also not puzzled. There are honest people, who keep their promises, including promises to pay back loans, and there are dishonest people who do not. If Bill trusts Alan's honesty, as he should if Alan has a reputation of being honest in the past, he will make the loan and sensibly expect repayment barring unforeseen contingencies. If Alan is honest, he will consider the past action of having promised as reason enough to repay the loan.

The economist, as part-time layman himself, may start from this sort of explanation, but he is not, or should not, be satisfied. Why should his past action of having promised to repay be reason enough, or reason at all, for Alan to repay the loan? Does this not contradict the adage that bygones are bygones, the teleological or forward-looking perspective both used by economists and attributed to rational economic agents? It would also seem somewhat ad hoc to put broken promises as an argument in some agents' utility functions. He then takes heart at the observation that Bill is more likely to trust Alan's honesty if he knows that Alan has repaid debts and kept promises in general in the past. This then is the incentive for Alan to keep his promise. Alan's repaying or failing to repay this loan will influence others' expectations of the likelihood that Alan will repay loans in the future, and hence influence the likeli-

hood that he will be able to borrow again in the future. Justice, or the keeping of contracts, is rational because it pays to be believed just.

I believe this explanation to be essentially correct. It may turn out to be inadequate to explain all empirical instances of promise keeping, but we should not resort to other explanations requiring more drastic revision of economic theory without first exploring the possibilities of this explanation. Even if we should eventually resort to explanations involving respect for promises for their own sake, with no thought of the consequences of being seen to keep promises, we might still seek to explain the origin of the practice which provides an object for such attitudes without invoking the attitude itself.

This essay simply explores in greater detail the argument that it is rational to keep promises in order to maintain a reputation for doing so. Of particular interest is the explanation of how and why current actions of keeping or failing to keep promises might influence others' beliefs that one will do so in future.

The "promising game" is a simple way of characterizing a wide class of games in which there arises the problem of the time-inconsistency of optimal plans under rational expectations. In all such games, agents would like to make credible commitments on their future moves in advance of making those moves. They would like to do so in order to influence others' expectations of their future moves. I present a very simple version of such a game which shows that the problem of time-inconsistency is quite general. The exceptions are the special cases. I then consider both finitely and infinitely repeated versions of the same game to explore the possibility of reputation. Contrary to what might be supposed, complete and perfect information can make it harder, not easier, to prove that reputation exists in equilibrium. To show this, we examine

the results of Milgrom and Roberts (1980) and Kreps and Wilson (1981) who prove the existence of reputation in finitely repeated games under imperfect information. What explains their results is that "amoral" players find it profitable to mimic the behaviour of players who are constrained to act as if they kept their promises, when other players cannot distinguish the two types a priori.

Under complete and perfect information, the problem is more complex. If the equilibrium outcome of the promising game is predictable, then the promisee can predict a priori whether or not the promiser will keep his future promises, and the promiser knows this. But then it seems that the promiser has no incentive to keep his current promises, for he knows that the expectations of the promisee are determined a priori. When the promiser decides whether to keep or break his promise, he considers the consequences of each of his two options for the beliefs and expectations of the promisee but one of the two possible actions must be a disequilbrium or irrational action, and the promisee would know this. What can the promisee believe when he observes an action he knows to be irrational? It is this fundamental incompleteness or indeterminacy of teleological modes of action and belief which opens the door for the rationality of non-teleological, "deontological" or rule-following modes of action, and the rationality of believing that agents can be committed to following rules such as the keeping of promises.

II. The Concept of Time-Inconsistency

I define a "method" of choosing a plan of action to be a mapping from an agent's preferences and beliefs into a plan of present and future actions. Thus to say that an agent follows a specific method means that he formulates a plan of action based, in some specified way, on his preferences and beliefs. By "plan of action" I mean also to include contingent actions, whereby a single plan of action may include several alternative actions, each to be performed contingent on the agent's learning particular information in the future.

An agent's method is said to suffer from "time-inconsistency" if the action he plans at time t to perform at time s differs from the action he plans at time t + i to perform at time s. An agent whose method suffers from time-inconsistency may plan today to do something tomorrow which differs from what he actually does when tomorrow arrives. Since I have defined a plan of action to include contingent actions time-inconsistency cannot be merely a facet of the receipt of new information. For example, an agent who plans to go to the beach tomorrow if the weather is fine, but when tomorrow arrives stays at home because it is raining, does not, at least on these grounds, follow a time-inconsistent method. If, on the other hand, he plans to go to the beach tomorrow come what may, but still ends up staying at home, then he does follow a time-inconsistent method.

It is important to notice that time-inconsistency pertains to the method of choosing a plan of action, rather than the plan of action itself. We need at least two plans, formulated at different times, for the question of time-consistency to arise. Therefore it is the method adopted to formulate those successive plans which may be time-consistent or time-inconsistent.

Obviously, if we place no restrictions on the set of possible methods, than any method chosen at random may turn out, as often as not, to be time-inconsistent. Our interest is not with planning per se, but with rational planning. We wish to examine the implications of the time-inconsistency problem for utility-maximizing methods of choosing plans. To anticipate our conclusion somewhat, it will turn out that, like time-consistency, rationality pertains to the method of choosing plans of action, and not to plans of actions themselves.

III. Time-Inconsistency of Preference

The first example of the time-inconsistency problem, due originally to Strotz (1955), arises from the nature of an agent's inter-temporal preferences. We can consider an agents preferences as a ranking over temporal sequences of states of the world. A temporal sequence is simply a dated list of states of the world, with one state for each and every relevant period of time. In order to check for time-inconsistency of an agent's preferences, we compare his ranking of sequences at two points in time, t and t + i, where all sequences within a given comparison set are identical up to t + i. If, within any comparison set,[1] the agent's preference rankings of sequences differ between times t and t + i, then the agent is said to have time-inconsistent preferences.

What is the rational method of formulating a plan of action for an agent with time-inconsistent preferences who knows how his preferences will evolve over time? Suppose an agent were simply to choose a plan of action which, if adhered to, would maximize his current preferences subject to his beliefs about any external constraints upon his choice, and implemented the current portion of that plan. This method is obviously vulnerable to time-inconsistency, for his future preferences, and hence future plans, may differ from today's. It is also obvious that such a method would be irrational, since it is based on the assumption that he will adhere to his current plan in the future, whereas he knows that his future preferences may dictate a deviation from the plan. As an alternative to this "naïve" method, a "sophisticated" method has been proposed, by, for example, Phelps and Pollack (1968). Following the sophisticated method, the agent predicts his own future action as a function of his future preferences and constraints. He then chooses his

current plan to maximize his current preferences subject to his perceived constraints, including the constraint of his own predicted future actions. Insofar as his future constraints or preferences, and hence future actions, depend on his current action, this sophisticated method will lead to a sequence of actions which yield him higher expected utility, according to his current preferences, than would the naive method. Furthermore, the sophisticated method is time-consistent in the sense that his currently predicted future actions coincide with his future choices. Notice, however, that the agent does not plan his future actions in any meaningful sense. He simply predicts them.

The case of time-inconsistent preferences is not central to this paper. It is examined here in order to distinguish it from the problem of time-inconsistency under rational expectations, and in order to gain some insight into the problem of time-inconsistency in general.

The essence of the sophisticated method of an agent with time-inconsistent preferences is that he acts towards the agent he will be tomorrow as if acting towards another person. In a two-period problem, the solution is formally equivalent to the Stackleberg leader-follower equilibrium for a duopoly. The current agent constructs the future agent's reaction function, and picks a point on it to maximize his own utlity, which depends on his own action and that of the future agent. The asymmetry between the players' strategies is simply due to the fact that the current agent's action, once past, cannot be changed.

It should come as no surprise that the solution for an agent with time-inconsistent preferences is the same as the solution for a game involving a sequence of different agents. In economic theory, an agent is a conjoined set of preferences and beliefs. If two agents have the same beliefs and strictly identical preferences (the same ranking over

states of the world), then they can be treated as if they were one and the same agent. If one agent has different or changing tastes over time, as in the case of time-inconsistent preferences, he can be treated as if he were a sequence of different agents. In other words, the agent with time- inconsistent preferences has no trans-temporal identity. He cannot choose or decide today, in any meaningful sense, what he will do tomorrow. He can only predict, and act now to influence his future choice.

IV. Time-Inconsistency under Rational Expectations

In a game against nature, the problem of time-inconsistency can only arise if the agent has time-inconsistent preferences. In a game against other agents, however, the problem of time-inconsistency of optimal plans can arise even if, as we shall henceforth assume, agents have time-consistent preferences. The problem arises because one agent's utility may depend on the action performed by another agent, which in turn may depend on the latter's expectation of the former agent's future actions. If that expectation is formed rationally, and on the basis of full information, it will coincide with the former agent's planned action. However, the plan chosen by the former agent ex ante, when he can apparently influence the other's expectation, may differ from the plan chosen ex post, when the other's expectation and action are predetermined.

The problem of the time-inconsistency of optimal plans, recognized as such, was first introduced to economics by Kydland and Prescott (1977), and used to interpret the distinction between rules and discretion in the context of economic policymaking. They show that adherence to a policy rule can outperform discretionary policymaking even under conditions of full information, because the two methods have different implications for agents' expectations of future policy actions. Kydland and Prescott, however, do not consider the credibility of the policymaker's decision to adhere to a rule. They simply assume, without argument, that agents will know whether a policy rule or discretionary policy will be followed.

Though Kydland and Prescott (1977) is considered to be the seminal work in this area in economics, essentially the same problem also appears, and appears earlier, in other disciplines. Many of the topics

covered in Schelling (1960) consider the rationality of, and the possibilities for, committing oneself in advance to a course of action which would not otherwise be a rational course of action ex post. Precommitment pays, when it does pay, because it influences to one's advantage the expectations of others concerning one's future actions. Agents can precommit themselves by taking current actions which conspicuously alter their future options or payoffs, to change the action it becomes rational to perform ex post, and hence rational to expect ex ante. Schelling also mentions the possibility that an agent might act as if he were thus committed, not because he in fact is, but in order to develop a reputation for acting thus and so lead others to expect that he will act simlarly in similar future games. However, a reputation cannot be created unilaterally. Is it rational, as Schelling (ibid., p. 30) contends, for agents to believe that a reputation will be maintained simply because it is rational to maintain others' beliefs that it will be maintained?

In formal game theory, Selten (1975), as described by Harsanyi (1976), showed that what appear to be equilibrium strategies for games in normal form (with the time-sequence of moves suppressed) cannot be equilibrium strategies when the game is described in extensive form (with the time-sequence of moves revealed), for they involve threats which are not credible, in the sense that it would not be rational to carry out that threat once the action one sought to deter had already taken place, and this is known by other players who are thus not deterred. Selten proposes that we restrict our attention to perfect equilibrium strategies - strategies which are equilibrium strategies in all branches of the game, whether those branches would be reached in equilibrium or not.

In Moral Philosophy, Hodgson (1967) uses what is essentially the time-inconsistency problem to point out a fundamental inconsistency in act- utilitarianism, which for our purposes can be interpreted as the ethical doctrine which maintains that an action is good if and only if the consequences of that action are good. He contrasts act-with rule-utilitarianism which (again for our purposes) can be interpreted as the ethical doctrine which maintains that an action is good if and only if it conforms to a good rule, and that a rule is good if and only if the consequences of following that rule are good.

Hodgson's central argument against act-utilitarianism is that:

> "...certain good consequences depend on the existence of expectations of actions, and that, under certain conditions as to knowledge and rationality, an agent's avowedly acting upon the act-utilitarian principle could preclude such expectations and such good consequences."[2] (Ibid., p. 85, italics in original).

Hodgon's argument in favour of rule- vs. act-utilitarianism thus corresponds to Kydland and Prescott's argument in favour of rules vs. discretionary policy. Of particular interest in Hodgson's argument is his recognition of the self-referential quality of the inconsistency in act-utilitarianism, that when act-utilitarianism is judged on act-utilitarian grounds, the principle rejects itself. He also argues, contra Schelling, that the paradox does not disappear even when the game is repeated many times - that the "bootstraps" argument for the rationality of maintaining a reputation is invalid, by appeal to the infinite regress involved in such an argument (ibid., p. 87). This argument, he notes, has profound implications for the institutions of promising,

language, and law, which cannot logically exist under common knowledge of act-utilitarian behaviour.

V. The Single Game

Below we give an example of the problem of time-inconsistency of optimal plans under rational expectations, which is a simplified version of the example given in Kydland and Prescott (1977), but which retains the essence of the problem.

Consider a simple, two person, two period game. Player A's utility is a function of player B's action in period one, y, and of A's action in period two, x. B's action in period one is a function of B's expectation in period one of A's action in period two, x^e. (For simplicity we suppress B's choice problem in order to focus on A's choice problem. In the single game this is of no consequence, but it does preclude B from strategic behaviour in the context of repeated games. We will return to this point later.) We will assume that x and y are continuous variables, and that all relevant derivatives exist.

$$U_A = U(y,x) \tag{1}$$

$$y = y(x^e) \tag{2}$$

To complete the description of the game we must specify the information available to the players. We will assume that both players know everything that we know - that equations one and two are "common knowledge" between A and B.

Definition: It is "common knowledge" between A and B that P if and only if:

(i) A and B both know that P and,

(ii) A and B both know that it is common knowledge

between A and B that P.

That the above definintion is self-referential does not imply that it is illegitimate. It should be interpreted as an algorithm which can recursively generate the infinite set of sentences of the form "A knows that B knows....that A knows that P" which together constitute the definition of common knowledge. I contend however, that the algorithmic definition is not merely a convenient shorthand for the extensional definition, but that it models the process of inference that occurs when two rational agents recognize each other as rational agents. To ascribe rationality to a being - to recognize him as constituting a person - involves not merely an ascription of preferences and beliefs to that being (as may be ascribed to intelligent animals) but involves also a recognition that he in turn may recognize one's own rationality, and hence recognize that one recognizes his rationality....etc. The possibility of common knowledge is an essential component of intersubjectivity - of the interrelation between one fully rational agent and another. It is a prerequisite of their being said to share the same society.

The assumption of common knowledge completes our specification of the game. We now consider possible solutions.

In period one, A chooses a plan of action which, if adhered to in period two, would maximize his utility subject to the constraint of equation two. His optimal plan depends on the way in which B forms his expectation of A's future action. Suppose initially that B's expectation of A's future action is identically equal to A's currently planned future action. Substituting equation two into A's utility function, and replacing x^e with x, A's optimal plan is such that

$$\underset{x}{\text{Max}}\ U_A = U(x, y(x)) \qquad (3)$$

A's optimal plan is such that

$$\frac{\partial U}{\partial x} + \frac{\partial U}{\partial y}\frac{\partial y}{\partial x} = 0 \qquad (4)$$

Solving equations two and four simultaneously yields the "deontological" equilibrium $\{x^D, y^D\}$, which we will asume to exist and be unique. This solution is formally equivalent to the Stackleberg solution in duopoly, with A the "leader" and B the "follower". A picks a point on B's reaction function which maximizes his utility. In this solution it is common knowledge between A and B that A is following the "deontological" method, he acts as if he were constrained to follow that plan of action which would have been optimal ex ante had it been common knowledge that he would act as if he were constrained. It is as if A can choose B's expectation x^e, setting it at whatever level he desires, subject only to the constraint that his action must subsequently validate B's expectation. Conversely, it is his commitment to validate B's expectation, his ability to "promise," which enables him to choose B's expectation.

In this game, however, there exists no way for A thus to commit himself in advance, nor to communicate that commitment to B. When period two arrives, A may reformulate his optimal plan, this time taking B's action as predetermined and hence parametric to his choice. Thus his second period optimal plan is given by:

$$\frac{\partial U}{\partial x} = 0 \qquad (5)$$

Realizing that his action will be parametric to A's choice, B forms his expectation by solving equations two and five simultaneously, which

yields the "teleological" equilibrium $\{x^T, y^T\}$, which we assume to exist and to be unique. Formally, this equilibrium corresonds to the Cournot equilibrium in duopoly; both players are "followers." In this equilibrium, it is common knowlede between both players that A follows the "teleological" method; he reformulates his optimal plan in each period and implements the current portion of that plan.

Except in the special case where the second term in equation four is equal to zero in equilibrium, which implies either that $\partial U/\partial y = 0$ or $\partial y/\partial x = 0$, the deontological and teleological solutions will differ.

Which method, the teleological or deontological, is the rational method for A to follow? Which method yields him higher utility? The answer depends on the manner in which B forms his expectation of A's future action. In order to form an expectation of A's future action, B must make an assumption about the method adopted by A, whether it be the deontological method of equation four, or the teleological method of equation five. Which method is rational for A to adopt, however, does not depend on which method B attributes to A. For any <u>given</u> method attributed by B to A, the teleological method always yields higher or equal utility compared to the deontological method. Rather, which method is rational depends on whether <u>the method attributed by B to A depends on the method actually adopted by A</u>. It pays to adopt the deontological method if and only if one's doing so is a necessary and sufficient condition for one's being believed to have done so.

To see this, first suppose that the method assumed by B is independent of the method actually adopted by A. Clearly the teleological method now dominates, for B's expectation and hence B's action is indeed independent of the method actually chosen, and so adopting the deontological method merely places the additional constraint of equation four

on A's action. Since B's action is parametric, it is rational to treat it as such as in equation five. Now suppose instead that the method assumed by B is the same as the method chosen by A. A can now choose between the teleological equilibrium $\{x^T, y^T\}$ and the deontological equilibrium $\{x^D, y^D\}$. Clearly, choosing the deontological method, and hence the deontological equlibrium, now yields an equal or higher level of utility, for by adopting the deontological method, A can influence B's expectation, and hence B's action, to his advantage. Formally, it pays to be a Stackleberg leader rather than a Cournot follower. By adopting the deontological method A can pick any point on B's reaction function. By adopting the teleological method A must pick the point where B's reaction function crosses his own reaction function.

It pays to adopt the deontological method if and only if one's doing so leads others to believe that one does so. The problem is; how is it possible thus to influence the expectations of others when they know that, when tomorrow arrives, the constraints of today's optimal plan will be purely imaginary? A would like to commit himself in advance to perform the action implied by the deontological method if by doing so he can convince B that he is thus committed. In the world as we know it, there exist ways in which an agent can conspicuously modify his payoffs to future actions, and hence influence the action it would rational for him to choose in the future, in order similarly to influence the action that others expect him to choose in the future. Such devices may be mechanistic, as when a pilot for a passenger airplane refrains from wearing a parachute, or legalistic, as when an agent signs a contract with penalties for non-compliance. However, such devices are not always available, and if they are available, are not without cost. Moreover, such devices are not solutions to the problem given above. To invoke such devices is

not to solve the game but to replace it with a different game which is like the first but with additional branches. That extended game is then solved sequentially to find Selten's perfect equilibrium which is the teleological equilibrium of the extended game. That equilibrium may look like the deontological equilibrium in the original game, but it cannot be the same equilibrium since it is the equilibrium of a different game. Similarly, a cooperative game, wherein it is assumed that agents can make binding commitments before play starts, is simply an incompletely specified non-cooperative game. To invoke such devices is not to answer, but to change the question.

The structure of our game does not include the use of devices to commit oneself in advance. In period two B's action is predetermined. A's choice of action thus cannot influence B's action, so he rationally chooses his action according to the teleological method of equation five. B anticipates this, and forms his expectation accordingly. The players are resigned to the teleological equilibrium, where the teleological method is both adopted and expected, because there is no way that A could convince B of his decision to adopt the deontological method.

VI. The Finitely-Repeated Game

In the preceeding section we assumed it to be common knowledge between both players that the game would be played once only. On the basis of this assumption we concluded that the teleological equilibrium $\{x^T, y^T\}$ was the solution to the game. Would our conclusion be modified if we introduced the possibility of repeated plays of the same game? Would it now be in the interest of player A to adopt the deontological method in order to convince B that he will play the same strategy in future games? Would it be rational for B to be thus convinced?

Let us assume that it is common knowledge between both players that the game will be repeated n times, where n is some finite number. It can be seen, given our assumptions, that the equilibrium in this finitely extended game must correspond to that of the single game. This result is known as the Selten chain-store paradox following Selten (1978), which is a close cousin of the well-known "hangman" paradox.

Consider the last game. Since the outcomes of all previous games are predetermined, and since there exist no future games which the outcome of this game can influence, this last game is exactly the same as the single game considered in the previous section. Thus it is common knowledge between both players that the outcome of the last game will be the teleological equilibrium. Consider now the penultimate game. Since the outcome of the next game is known in advance to be the teleological equilibrium, it cannot be influenced by the outcome of the penultimate game, which therefore is also exactly like the single game, and will result in the teleological equilibrium. Our argument proceeds by mathematical induction back to the very first game, and entails that the outcome of any extended game which is commonly known to be of a certain,

finite length, must correspond to the teleological equilibrium in the single game.

What drives this argument is the proposition that the outcomes of future games can be known in advance. Information on the outcomes of preceeding games is not necessary to predict the outcomes of future games. Since A knows that B will predict the outcomes of future games independently of A's current choice, A's current choice cannot influence B's future expectations, and so cannot influence the outcomes of future games. Thus each game becomes separated from the predetermined outcomes of previous games, and the independently-predictable outcomes of future games. What researchers in this field have generally failed to perceive, however, is that a very similar argument can be applied, not only to the finite, but also to the infinite game.

VII. The Infinitely-Repeated Game

In the previous section we used the "hangman" argument of the Selten chain-store paradox to show that if the game were to be repeated a finite number of times, and if this were common knowledge between both players, then the outcome of each game in the finitely-repeated game must correspond to the teleological equilibrium $\{x^T, y^T\}$ in the single game.

The "hangman" argument starts by showing that the outcome of the final game must be the teleological equilibrium, and then proceeds backwards by induction to show that all preceeding games must have the same outcome. It might be thought, therefore, that if we somehow modify our assumption of common knowledge of the finite number of repetitions, then the "hangman" argument no longer applies. We could assume instead that it is common knowledge that the game will be repeated an infinite number of times. More weakly, we could simply deny that it is common knowledge that when the final game arrives, A will know that it is the final game. Reasoning such as this might lead researchers into arguing that in the infinitely-repeated supergame, since the "hangman" argument does not apply, there is no problem in positing the deontological equilibrium as one possible outcome, as Milgrom and Roberts (1980) argue, or as the only possible outcome, as Kurz (1977) argues.

Strictly speaking, of course, the "hangman" argument can not be applied to the infinitely-repeated supergame, for the former starts from the final game, which does not exist in the latter. However, what lies behind the "hangman" argument is not so much that the next game is the last, but the idea that the outcomes of the next and future games are predictable independently of knowledge of the outcomes of previous games.

Once this is recognised, the "hangman" argument, or rather a variant of it, can equally well be applied to infinite as it can to finite games.

Let us suppose that it is common knowledge between both players that the game of section five will be repeated an infinite number of times. Can we predict the outcome of this supergame, and if so, what outcome will we predict?

Let us assume that our knowledge of the supergame is "complete". By this I mean that the description of the supergame is sufficient information to predict the outcome of each and every game in the supergame up to an irreducibly random and unknowable element (this last clause is to permit the possibility of mixed, or random strategies). In particular, we can, at any point in time, predict the outcomes of all future games without needing to have observed the outcomes of any previous games. From our assumption that the description of the supergame is common knowledge between both players, however, any proposition we know or can deduce about the game is also known by both players. If, as we assume, our knowledge of the supergame is "complete", then so is that of the players. In particular, if we can predict the outcomes of all future games a priori, then B too can predict the outcomes of all future games (up to an irreducibly random element) without needing to have observed the outcomes of previous games. And if we know that B can predict thus, then A knows this also. Now A's only motive for playing X^D in the current game is if his doing so might influence B's expectation of A's play in future games, but if B already knows how A will play in future games (again, up to an irreducibly random element) his expectation cannot thus be influenced, and this is known to A. Thus A has no motive to play X^D, and, taking B's expectation as predetermined a priori, rationally plays X^T, knowing that B knows this and will expect X^T and will reply with y^T.

To summarise; assuming that a complete description of the infinitely-repeated supergame is common knowledge between both players, we concluded that <u>if</u> the description of the game is sufficient to predict the outcome of the supergame, then the only possible outcome we can predict is a repetition of the teleological equilibrium $\{x^T, y^T\}$ of the single game. Alternatively, if some other outcome is possible then we cannot predict that outcome a priori.

The intuition behind this argument is as follows; A knows that B will only revise his expectation of A's future actions if he receives additional information, but it seems that B already knows everything that can be known on the determinants of A's rational choice. B knows A's preferences and all A's beliefs, and A knows that B knows this. What additional information could there possibly be which B might find useful in predicting A's future actions? If there is none, then A can give him none.

VIII. Interpretation of Preceding Results

In the above section, assuming common knowledge, we concluded that, even if the game is infinite, if the outcome of the game can be predicted a priori then the outcome we predict must be the teleological equilibrium $\{x^T, y^T\}$. This conclusion is paradoxical in (at least) two ways.

Firstly, the conclusion is paradoxical because it seems to contradict common experience. The game can be interpreted as a "promising game" whereby in period one A promises to perform x^D in period two, in order to get B to perform y^D in period one. When period two arrives, however, A has an incentive to break his promise and to perform an action other than x^D. Even though promises are costly to keep, we normally argue that A has an incentive to keep his promise, and hence B has reason to believe he will keep his promise, because B's belief that A will keep his promises in the future depends, at least in part, on whether A has kept or broken his promise in the past. What the argument in the above section implies, however, is that B's beliefs concerning A's future actions are determined a priori independently of A's past actions, and that since this is common knowledge between A and B, A has no motive ever to keep his promises, qua promises, and that B knows this. A's keeping or breaking his promises in past games gives B no information on the determinants of A's rational choice in future games.

What is surprising, from our analysis, is not that promises are sometimes broken, but that they are ever kept at all, that they are ever expected to be kept, that they are ever made, and that the very word should appear in our language. The same applies not only when the keeping of a promise is beneficial to the other, as in repaying a loan, but also when it is harmful, as when the deterrent of punishment is used

to enforce a law. It applies whenever the performance of an act is, in itself, costly to the performing agent, but is beneficial only insofar as it establishes a reputation, an expectation by others that acts of that type will be performed in future. What our argument seems to imply is that it is impossible thus to establish a reputation, and therefore irrational ever to try. It is not enough to counter that promises are sometimes conspicuously enforceable by the threat of punishment by third parties, for sometimes they palpably are not. Even where they are, to invoke enforcement by third parties is merely to pose the same problem in a different place. How do those third parties establish a reputation for enforcing the promises of others? How can the emission of sound waves, or the production of ink marks on a piece of paper, except trivially, change the world? How can they ostensibly alter the payoff structure which determines rational choice?

Secondly, the conclusion is paradoxical because it seems to imply that a non-rational player, or mechanism, could attain the ends of a rational player better than could the latter himself. Let us replace our rational agent A with a mechanism A^1, which simply repeats x^D every game. We will assume that B knows he is confronted with a mechanism, but it matters little whether B knows the structure of that mechanism. If B knows the structure, he will expect x^D, and the outcome is immediately $\{x^D, y^D\}$. If B does not know the structure, he will observe X^D repeated, and will eventually revise his expectations towards x^D, and the outcome will soon become $\{x^D, y^D\}$ similarly. If B did not thus revise his expectation, he would make consistent mistakes, which would seem irrational. Thus the mechanism could attain an outcome which would be preferred by the rational agent A to that which he himself could attain.

Why cannot the rational agent A attain the same outcome as the non-rational mechanism? Why cannot he simply act as if he were such a mechanism? Any difference must lie in the nature of the constraints they face. In particular, the rational agent in the argument in the preceding section is constrained by B's prior knowledge of the determinants of his rational choice. The mechanism is not thus constrained.

B has prior knowledge of the determinants of A's rational choice. Does this mean that A's choice is pre-determined? Obviously not, for an agent with free will can always choose any action whatsoever, rational or irrational, but why should he ever want to choose an irrational action? The idea sounds like a logical contradiction, and yet an agent could want to do so if he intends thereby to destroy other agents' ability to predict his action on the basis of their knowledge of those particular determinants. An action which is irrational at one level of understanding can become rational at a higher level of understanding insofar as it destroys that primary level of understanding. In the game at hand, A does indeed have reason to seek to destroy B's ability to predict A's action on the basis of his a priori knowledge of A's preferences and beliefs. It is the constraint of B's forming his expectations in this way that prevents A from attaining the outcome that can be attained by his machine - counterpart A^1.

IX. Imperfect Information

The only motive that player A could have for playing x^D is if his doing so should lead player B to expect him to play x^D in future games. From the argument in preceding sections, however, it would seem that player A is unable to attain the deontological equilibrium $\{x^D, y^D\}$ because his current play does not influence B's expectation of his play in future games. This is so because, under the assumed conditions of common knowledge, B already knows a priori all the determinants of A's rational choice - his preferences and beliefs - and so A can give B no additional information which might lead B to revise his expectation.

In recent contributions, Milgrom and Roberts (1980), and Kreps and Wilson (1981) have sought to reverse this conclusion by modifying the common knowledge assumption. The context of their discussions is the problem faced by a monopolist in seeking to deter, by predation, the entry to his market of a succession of firms. The established monopolist always prefers to maintain his monopoly position. Faced with the fact of entry, however, he always prefers to share the market rather than to engage in mutually costly predation, except insofar as his decision to prey this period might deter future potential entrants. The potential entrant, on the other hand, prefers to enter if he knows the monopolist will share the market, and prefers not to enter if he knows the monopolist will prey. Thus the monopolist's strategy "prey if entry occurs" corresponds to the deontological method, while the strategy "share if entry occurs" corresponds to the teleological method.

In the finitely-repeated game, where the monopolist faces a finite number of potential entrants to a finite sequence of markets, and under conditions of common knowledge of preferences and beliefs, Milgrom and

Roberts, and Kreps and Wilson, (hereafter M-R and K-W) recognise that the "hangman" reasoning of the Selten chain-store paradox entails that the only equilibrium outcome is where the potential entrants always enter, and the established monopolist never preys - the teleological equilibrium.

In order to avoid this conclusion, M-R and K-W modify the assumption of common knowledge. They introduce the possibility of a "strong" monopolist, who, either because he has a different payoff structure, or else because he is simply not rational, will always prey when faced with entry, regardless of any effect this might or might not have on the decisions of future potential entrants. The monopolist knows with certainty whether he is "strong" or "weak", but potential entrants do not know this with certainty a priori. Instead they entertain a small, but non-zero probability that the monopolist is strong. From this point on, the game as described, including the history of moves, is common knowledge between all players.

Given this small asymmetry of information it is easy to prove that sharing markets is not necessarily the equilibrium strategy for a weak monopolist (i.e. it is not his equilibrium strategy under all parameter values). This can be proved by reductio ad absurdam.

Assume that the equilibrium strategy for a weak monopolist is to share if entry occurs. If this is the equilibrium strategy for a weak monopolist, then potential entrants will know it to be the equilibrium strategy. Suppose that entry occurs in one market. If entry is met with predation, then potential entrants will know with certainty that the monopolist is strong, since weak monopolists, by assumption, never prey, and so will not enter in any future market. Conversely, if entry is met with sharing, it will become henceforth common knowledge that the monopo-

list is weak, and potential entrants will always enter in future. Given that potential entrants form their beliefs in this way, the weak monopolist, faced with entry, can either prey this time and have sole control of all future markets, or else share and be faced with entry in all future markets. For some discount rates and payoff structures the present value of taking the former choice will exceed that of the latter, and the weak monopolist's optimal strategy will be to prey if entry occurs. This contradicts our initial assumption that his equilibrium strategy is necessarily to share.

Could it ever be an equilibrium strategy for a weak monopolist to prey in order to promote the expectation that he will prey in future? M-R and K-W argue that it is. In the finitely-repeated game the weak monopolist will generally prey in early stages of the game, share in the later stages, and perhaps adopt a mixed strategy in between. For our purposes however, it is sufficient, and simpler, to consider the equilibrium strategy of a weak monopolist in an infinitely-repeated game.

To examine the reasoning behind the result obtained by M-R and K-W, we first posit that the equilibrium strategy for the weak monopolist is indeed to prey if entry occurs. We then construct the best reply to this strategy on the part of potential entrants, which, of course, is never to enter. Given this best reply, we then check to see if the weak monopolist has any incentive to depart from the strategy we posited initially, both at equilibrium and at <u>disequilibrium</u> points in the game tree. (The latter is to ensure that the threat of predation is credible, or a perfect equilibrium strategy in the sense of Selten.) If the monopolist has no incentive to depart from his strategy, then it is indeed an equilibrium strategy.

Suppose that entry occurs in one market. If the monopolist maintains his strategy, and preys, he assumes equilibrium outcomes in all future games, which means he will never again face entry. If the monopolist departs from his strategy, and fails to prey, K-W argue, then he knows it will become common knowledge between himself and all potential entrants that he is in fact weak, since strong monopolists never fail to prey. He also knows that if it becomes common knowledge that he is weak, then the equilibrium in all future games must be entry and sharing. Given a low enough discount rate, the weak monopolist will prefer the former to the latter sequence of outcomes, and so will not choose to depart from the strategy we posited, which shows that it is indeed his equilibrium strategy.

I find this argument to be problematic. Notice that what sustains the equilibrium is the assumed interpretation that potential entrants would place on a disequilibrium move by the weak monopolist. It is assumed that the potential entrants would interpret failure to prey as proof that the monopolist was weak, for a strong monopolist would not, or could not, fail to prey, despite the assertion that failing to prey is not a rational strategy on the part of the weak monopolist either.

If the entrant knows that the equilibrium strategy for both weak and strong monopolist is always to prey, then observing failure to prey confronts him with a logical contradiction. One way to avoid such contradiction is to invoke the "trembling hand", as in Selten (1975). We suppose that each player will make "mistakes" with small but non-zero probability. A player faced with a disequilibrium move by another player can then interpret that move as a "mistake". If we follow this route, however, the argument used by K-W must collapse. By symmetry we should assume that both weak and strong monopolists have the same probability of

making a mistake. If this is assumed, then observing a mistake would give the potential entrants no information on whether the monopolist is likely to be weak or strong - they would not modify their prior probabilities as a result of observing failure to prey. Knowing this, the weak monopolist would choose not to prey, which contradicts our initial assumption that his equilibrium strategy is to prey if entry occurs. The argument used by K-W rests on an arbitrary implicit assumption that the weak monopolist makes mistakes with greater probability than does the strong.

The model put forward by M-R contains a slight difference which enables them to avoid the above problem. They posit a third type of monopolist, we might call him a "wimp", who, for some reason or another, will always share if entry occurs, regardless of any effect this might have on future entry. The potential entrants can now interpret failure to prey as proof that the monopolist is a wimp, to whom their best reply is always to enter. The weak monopolist will then prey if entry occurs because failure to prey will lead potential entrants to conclude that he is a wimp. This leads us to the general observation that if we seek to avoid making arbitrary judgements about how players will respond to a move they know to be a disequilibrium move - about how they will interpret contradictory evidence - we must ensure that there are enough types of players so that every move (indeed, every possible sequence of moves) is an equilibrium move for at least one type of player.

X. Undecidability under Perfect Information

We are now in a position to compare the outcome of the game in section seven, where all knowledge was common knowledge, to the outcome of the game in section nine, where player B (the sequence of potential entrants) does not know with certainty the type of player A (the monopolist). It is obvious that the equilibrium where entry is deterred corresponds to the deontological equilibrium $\{x^D, y^D\}$, and that sharing corresponds to the teleological equilibrium $\{x^T, y^T\}$. In section seven we argued that the teleological equilibrium was the outcome, while in section nine we argued the contrary. Does this difference arise simply from relaxing the assumption of common knowledge? How general is the result of section nine?

In section seven we argued that the teleological equilibrium must be the outcome because, since B knows a priori all A's preferences and beliefs, A can give B no additional information which might lead B to revise his beliefs concerning A's rational action. Suppose, however, that A were to depart from his assumed equilibrium strategy. Suppose A were to play other than x^T? B would now be faced with a contradiction, a logical inconsistency, between his prior belief that A must play x^T, and the evidence that A does not play x^T in fact. It is now obvious that we previously assumed that, faced with such contradictory evidence, B would retain his prior beliefs and simply ignore that evidence. In effect, we implicitly assumed that B would interpret such a play as a "mistake", and would not modify his prior beliefs and expectation of future plays. If A knows that B will make this interpretation, then it is indeed a mistake for A to play other than x^T.

In section nine we assumed that if the weak monopolist were ever to

depart from any posited equilibrium strategy then potential entrants would interpret that as evidence concerning the monopolist's type. In modifying their beliefs concerning the monopolist's type, they modify their beliefs concerning the monopolist's future plays. The purpose of introducing additional arbitrary types of monopolist is simply to avoid confronting potential entrants with a logical contradiction in the event of a disequilibrium move on the part of the weak monopolist. Every move is an equilibrium move on the part of at least one type, so the potential entrant can always interpret any move without logical inconsistency, without having to interpret a move as a "mistake".

Introducing such additional types as the strong monopolist and wimp serves not so much to alter the outcome of the game but rather as a way to avoid having to answer a difficult question: How do agents react when they observe a move they know to be a disequilibrium move? It is no answer to say that such moves are never in fact observed. Any agent is free to make whatever move he chooses. In deciding which move to choose he contemplates, for all possible moves, what would happen if he chose that move. What would happen depends, in part, on how other agents will interpret that move. We cannot decide which is the equilibrium or rational move without consjdering how other agents would react if he were to take some other move.

In section seven we assumed, without argument, that B would interpret a move of x^D as a "mistake", or rather, we assumed that A would assume that B would interpret x^D as a mistake. Obviously this is not good enough. To interpret a move as a "mistake" is to fail to interpret it at all. It amounts to simply ignoring the fact that the move took place at all, because it contradicts B's prior beliefs. Yet if we reject this "interpretation", what interpretation is the correct interpretation?

I suggest that we simply do not know, and yet, paradoxically, our very ignorance of what constitutes the proper interpretation a priori can provide us with the means to reconcile prior beliefs of what constitutes equilibrium play with actual observed play. Any observed play can consistently be interpreted as providing evidence about the player's belief concerning how it is likely to be interpreted.

Suppose that A plays x^D. B knows that if A had believed that B would interpret this play as a mistake and not revise his expectations as a result, then a play of x^D would indeed be a mistake. B can maintain his hypothesis of A's rationality, can provide a consistent interpretation of A's move, only if he assumes that A believed that B would not ignore such a play (interpret it as a mistake) but consider it a rational, deliberate choice on the part of A. Furthermore, if B interprets A's play of x^D as rational, and as providing evidence of A's belief that B will interpret it as rational, then if B assumes that A will maintain this belief in future he will expect A to play x^D in future games also. Expecting x^D in future, B will reply with y, and A's play of x^D will indeed have been rational.

Suppose that A plays x^T. B knows that this would be a rational move on the part of A only if A believes that B would interpret A's move of x^D as a mistake. B therefore assumes that A believes that B would interpret x^D as a mistake. If B expects that A will maintain this belief in future games, he will expect x^T in future games, and will reply with y^T.

The only motive for A to play x^D is if he can thereby influence B's expectation of A's future plays. A's future rational plays depend on A's preferences and beliefs. Therefore A can influence B's expectation only if he can influence B's beliefs about A's preferences and beliefs. We had previously argued that since the structure of the game was common

knowledge between A and B there was no relevant information that A could give B which would modify B's expectation. As a result, we argued, A has no motive to play x^D. We have since seen, however, that knowledge of the structure of the game is insufficient to determine the equilibrium, to determine A's rational choice, without an additional hypothesis about how A expects B to interpret A's moves. This extra bit of relevant information is what A can communicate to B in making his move, and since he can thus influence B's expectation he does have a motive to play x^D. It is rational for A to play x^D provided that the present value of the outcome $\{x^D, y^D\}$ in all future games exceeds the present value of a single outcome $\{x^T, y^D\}$ and an outcome $\{x^T, Y^T\}$ for all games thereafter.

We can reformulate this argument in a slightly different way. Suppose we do not know whether it is rational to adopt the deontological or the teleological method. B therefore envisages playing against one of two possible types of agent: the type which believes it is rational to adopt the deontological method and the type which believes it is rational to adopt the teleological method. If B observes x^D he will learn that he is playing against the former type, and will expect x^D in future. If B observes x^T he will learn that he is playing against the latter type, and will expect x^T in future.

The link between A's past action and B's future expectation is essential if the deontological equilibrium $\{x^D, y^D\}$ is to be the outcome. This link is not provided by B's learning any new information about A's basic beliefs and preferences (beliefs about the structural parameters of the game and beliefs thereon), for by assumption these he already knows. What B learns of is A's decision to adopt the deontological method of action, and hence of A's belief in the rationality of that decision. A's belief in the rationality of adopting the deontological method cannot be

reduced to a logically equivalent statement about A's belief about the structural parameters of the game, for we have seen that the attempt to do so yields only an undecidability result for the rationality of an action.

Should A ever succumb to the temptation of departing from the deontological method, his doing so reveals only that he no longer believes it to be rational to maintain the deontological method. It is only the effect of this revelation on B's expectations which provides A with an incentive to maintain the deontological method. The only thing that prevents A from treating his past optimal plan as an irrelevant bygone is the effect of revealing that he considers his past optimal plans as irrelevant bygones.

XI. Conclusion

The concept of rational action that pertains to a game against agents with rational expectations is in general fundamentally different from the concept of rational action that pertains to a game against nature. This dictum, previously noted by Kydland and Prescott (1977), I have sought to elucidate in this essay. In a game against nature, to establish the rationality of an action it is sufficient merely to enumerate the consequences of all possible actions. Whichever action yields consequences of the highest expected utility is the rational action. In a game against agents with rational expectations, the consequences of any action depend, in part, on the effect of that action on the expectations of other agents. The effects of an action on the expectations of other agents in turn depends on whether they interpret that action as rational.

It is not enough to say that it is rational to choose whichever action maximises an agent's (expected) utility subject to his perceived constraints. In a game against agents with rational expectations, those constraints are not invariant to their perceptions of the method adopted to choose actions. Posed in this way, the problem is clearly related to Lucas' (1975) insight that the structure of an economy facing a policymaker may not be invariant to agents' beliefs about the policy regime in effect. We are simply exploring the implications of the "Lucas Problem" for optimal policymaking. The implication seems to be that one should consider the outcomes of all possible policy rules, assuming in each case that agents believe that policy rule will be followed, and choose whichever rule maximises the policymaker's objective function. If the policymaker can commit himself in advance, and be seen to do so, then this

policy rule is indeed the one to which he should rationally commit himself. Suppose, however, that such commitment, in the sense of conspicuously imposing sufficient external penalties on oneself for violating the rule, is not available as an option, or is not available without significant cost. Is it rational to act as if one were thus committed? Is following that policy rule a credible strategy?

The relevance of such questions is by no means restricted to macroeconomic policy. We punish criminals in the belief that punishment deters crime, yet the criminal act which is being punished is past and cannot be undone, and the act of punishment is costly to the punisher. We sometimes keep promises, even though it is costly to do so and the benefit one gains from making a promise has already been received and cannot be taken away. The only motive for acts of punishment and promise-keeping must be the effect of such actions, or failures to act, on the expectations of others. The maintenance of such practices, and the belief in their maintenance, are what constitute the very fabric of social order. Why are acts of punishing and promise-keeping rational? Why should such acts, or their opposites, effect agents' expectations?

Consider a world in which the basic parameters (such things as tastes and technology) are fixed and common knowledge to all agents. Each agent's expectations of actions of other agents depends only on those basic parameters. Each agent knows, therefore, that since he cannot change the basic parameters, nothing he can do can in any way affect the expectations of others. Thus punishing, keeping promises etc. are irrational, because such acts are rational only if they affect the expectations of others, but since the latter depend only on the basic parameters, they cannot thus be affected.

Milgrom and Roberts (1980) and Kreps and Wilson (1981) seek to escape this paradox by relaxing the assumption of common knowledge of the basic parameters. By introducing a few agents with exactly the right innate propensities to punish criminals, keep promises, and so forth, who cannot be distinguished a priori from "normal" agents, they can derive an equilibrium in which "normal" agents choose to mimic the punishing and promise-keeping behaviour of the "abnormal" agents.

I contend that if we seek a radical or fundamental explanation of the institution of the promise-keeping then we must reject this device of introducing "abnormal" types of agent. Postulating that agents have a taste for keeping promises would not constitute a non-vacuous explanation of why the institution of promising exists. Postulating a <u>probability</u> that agents have such tastes is but a small advance. It amounts to supposing that the mere uttering of the words "I promise..." might per se influence an agent's future behaviour. Now it may indeed be true that agents living in a society which contains the institution of promising may <u>develop</u> a sense of self-respect or of shame if promises are kept or broken, but an explanation of the institution which creates the object for such attitudes must not presuppose the existence of such attitudes. We are not born with a disposition to feel shame on taking a certain course of action consequent upon emitting certain sounds.

I propose an alternative solution. Introducing "abnormal" types of agent retains much of the adhocness of simply putting broken promises in agents' utility functions - postulating a taste for honesty or revenge. Moreover, it does not answer but merely avoids asking the question of what would happen if such abnormal types did not exist.

In order to be able to prove that a postulated equilibrium is indeed an equilibrium outcome it is necessary to be able to specify the

(expected) consequences of all possible actions available to an agent, including putatively disequilibrium actions. Those consequences depend on the effect of that action on the expectations of others, but if there is no type of agent for whom that action is a postulated equilibrium action, then it is impossible to predict how other agents might change their beliefs as a result of observing that action, and hence impossible to predict their reaction. What should one believe when one observes evidence which contradicts beliefs held with certainty? It is always an option, for an agent with free will, to contradict by his action one's knowledge of how he will act. If he has incentive thus to contradict one's knowledge one cannot know how he will act.

Faced with this formal undecidability of the consequences of actions, and hence the formal undecidability of equilibrium actions and expectations, agents can only look at past actions as a guide to future actions. Past acts of punishing and promise-keeping can only be interpreted as <u>decisions</u> to adopt the rules of punishing or of promise-keeping. Yet it is the formal undecidability of equilibrium expectations in a teleological world, and the consequent necessity of looking at the past as a guide to the future, in providing the link between current acts of punishing and promise-keeping and the expectation of future acts thereof, which provides the teleological incentive to punish and to keep promises. The belief that agents can follow rules provides the rationale for following rules, which provides the credibility of - the rationality of believing in the maintenance of - such rules. The formal undecidability of teleological equilibrium is what makes possible the deontological equilibrium, where agents follow rules and believe those rules will be followed.

The sort of reasons one can give for an act of punishing, promise-keeping or other instances of following a rule are fundamentally distinct from the sort of reasons one can give for other types of action. In the latter it is sufficient merely to point to the beneficence of the consequences of that act in itself. The reasons for an act of punishing and promise-keeping must look backwards at the past act to be punished or the past act of promising, at the past optimal plans of the agent. Sihgle acts of punishing and of promise-keeping are rational insofar as they are necessary to maintain the belief that a rational rule is being followed. The rule of punishing or promise keeping is rational insofar as the consequences of following such a rule and being believed to do so are beneficial. We cannot ignore the middle step. We cannot dispense with the concept of rules.[3]

I believe I have shown that it can be rational to act as if one is committed, and the corollary, that such commitment can be credible without external constraints. However, this by no means completes the analysis of such behaviour. Two unresolved problems in particular present themselves.

The first problem is that incentives for commitment (provided preferences are time-consistent) apply only in games of two or more persons, yet in this essay the choice-problem of the second agent (B) has been heavily suppressed. In general the second agent may have an incentive to commit himself also, perhaps to break or modify the commitment of the first. Potential entrants might wish to have a reputation for entering despite the threat of predation, criminals to commit crimes despite the threat of punishment. They might gain by committing themselves in order to break the commitment of the other. It is possible to set aside this problem by assuming the other agent to be a sequence of agents, as is

done by M-R and K-W, so that none can benefit himself by developing such a reputation. Similarly we could assume that there is a large number of anonymous criminals, so that all would free ride on the investment in reputation made by anyone. We cannot, however, resort to such devices in a general analysis of reputation.

The second unresolved problem, if we do not introduce a plethora of agent-types as in M-R and K-W, is the "hangman" paradox in the finitely-repeated game. It makes sense to assume that agents will decide the same in future as in past game-stages if all stages are the same. In the finitely-repeated game, however, future stages differ in that they have fewer future stages than do present stages. This does not mean, however, that reputation must be impossible in finitely-repeated games under common knowledge. It is still open to an agent to force a contradiction by making a "disequilibrium" move, and we cannot tell if it is a disequilibrium move unless we know how it will be interpreted. The problem with finite games is that the cost of maintaining a reputation remains constant through time, while the benefit declines through time to zero. At some stage the cost must exceed the benefit. The problem is not as general as it might seem, for making the end of the game uncertain, or allowing reputation to be a matter of degree might prevent the cost from ever exceeding the benefit, but it still remains as a problem.

Footnotes

* This is a revised version of part of an earlier essay "The Rationality of Deontological Methods," first written in July 1980. I would like to thank, without implication, Rosslyn Emmerson, Joel Fried, David Laidler, Tim Lane, Michael Parkin, Tom Rymes, and Ron Wintrobe for valuable comments on earlier drafts.

1. I have restricted the comparison sets of sequences in the above manner because in some sense, though in an uninteresting sense, any agent who does not live on memories alone wishes he had lead an extremely abstemious past in order to store up pleasures for the present and future. The more interesting case of time-inconsistent preferences, to which the above definition restricts attention, arises when an agent's current preferences concerning the future differ from his future preferences concerning the future.

2. The examples Hodgson uses to demonstrate his argument are singularly ill-chosen, however, in that he supposes all agents to be act-utilitarians, which means that they will all adopt the same ranking over states of the world. Such agents would never be tempted to break their promises, for the promiser and promisee would always agree on whether or not the promised action should be undertaken. Following rules is thus redundant for such agents, except in the very weak sense that they might need conventions to enable them to coordinate their choices between several equally optimal states (e.g., all drive on the same side of the road).

3. Rawls (1955) reaches essentially the same conclusion from an ethical perspective in his rule-utilitarian analysis of punishment. See also Harrod (1936).

References

Harrod, R.F. (1936), "Utilitarianism Revised," Mind, April.

Harsanyi, J.C. (1976), Essays on Ethics, Social Behaviour and Scientific Explanation, D. Reidel, Boston.

Hodgson, D.H. (1967), The Consequences of Utilitarism, Clarendon Press, Oxford

Kreps, D.M. and R. Wilson (1981), "Reputation and Imperfect Information," mimeo, Stanford University.

Kydland, F.E. and E.C. Prescott (1977), "Rules Rather than Discretion: The Inconsistency of Optimal Plans," Journal of Political Economy.

Kurz, M., (1977), "Altruistic Equilibrium" in B. Balassa and R. Nelson eds., Economic Progress, Private Values, and Public Policy: Essays in Honor of William Fellner, Amsterdam.

Lucas, R.E. (1975), "Econometric Policy Evaluation: A Critique", in The Phillips Curve and Labour Markets, eds. K. Brunner and A.M. Meltzer, North Holland.

Milgrom, P., and J. Roberts (1980), "Predation, Reputation, and Entry Deterrence," mimeo, Stanford University.

Phelps, E.S., and R.A. Pollack (1968), "On Second-Best National Saving and Game-Equilibrium Growth," Review of Economic Studies, April.

Rawls, J. (1955), "Two Concepts of Rules," Philosophical Review, v. 64 (January).

Schelling, T.S. (1960), The Strategy of Conflict, Cambridge, Harvard University Press.

Selten, R. (1975), "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," International Journal of Game Theory, Vol. 4.

Selten, R. (1978), "The Chain-Store Paradox," Theory and Decision, Vol. 9.

Strotz, R.M., (1955), "Myopia and Inconsistency in Dynamic Utility Maximisation," Review of Economic Studies.

Chapter Three

An Economic Explanation of the Nature and Existence of Social Institutions

I. Introduction

The world which presents itself to an economic agent is a world comprised of not merely of "physical facts" but also of "social facts". His actions depend not only on his beliefs about the height, width, mass, velocity, colour, temperature, etc. of objects and agents, but also on his beliefs about such facts as ownership, contractual obligations, and moneyness.

Unlike physical facts, the truth or falsity of social facts - such as "This land is owned by A", "This paper is money", and "A owes B ten dollars" - is not independent of agents' believing their truth or falsity. The physical facts are objectively given to agents in a way that the social facts are not. The very possibility of there being social facts presupposes the existence of social institutions - in this case of property, monetary exchange, and contract. No amount of observing the physical characteristics of the relevant objects or agents can suffice to establish the truth of a social fact without reference to the relevant social institution. We may look in vain for any physical characteristic of the area of land which constitutes its being owned by A.[1]

While any single individual takes the objectivity of social facts as given, this is not true for society as a whole. If no one believes in property, money, or debt, then property, money and debt do not, and cannot exist.

As is argued by Berger and Luckmann (1966), the social facts, and the social institutions which grant them their meaningfulness, are a socially constructed reality. Their existence is a facet of human action and belief. This does not mean that social institutions are constructed in the same sense that buildings are constructed. The building, as

bricks and mortar, has an ontological status which is logically independent of the actions and the beliefs which caused its existence. We can conceive of its existing independently of any human action or belief. The existence of a social institution like property, on the other hand, logically presupposes certain actions and beliefs on the part of the relevant agents. It would be logical nonsense to say that property rights existed if agents made no conceptual or behavioural distinction between "mine and thine". Social institutions are not a causal product of, but are identical to, certain types of human action and belief.

If social institutions are a facet of human action and belief, then an economic analysis of their existence should be possible. Understood as the consistent application of the methodological assumption of rational individual action and belief, the economic approach contains no internal scope restriction which would rule out such an application a priori. An economic analysis of social institutions is not, however, important merely to extend the proven scope of applicability of the economic approach. The "facts" assumed and explained by economic theories, such as endowments and exchanges, are for the most part social facts, and presuppose the existence of the relevant social institutions.

The absence of an explanation of the existence of social institutions leaves open the possibiity that economic theory, as applied to the analysis of action and belief within a given institutional framework, cannot be applied successfully to the analysis of action and belief towards that institutional framework. If that latter analysis were shown to be impossible, then the economic approach to human behaviour would be put in the embarrassing position of being a theory which, if generally true, would have no social facts to comprise its subject matter. Such doubt would be infectious, for how can we know that a big enought shift

in the demand for a good would result, not in a rise in price and quantity, but in a shift in the behaviour which constitutes the existence of property rights over that good - an erasure of the social facts implicit in the axes and curves of our model?

For these reasons an economic analysis of the existence of social institutions is important, but we should take care to understand what is meant by such an analysis.

By imposing ad hoc restrictions on agents' behaviour we could model the consequences of the existence of particular types of social institution, and ascertain (amongst other things) which type would be most efficient under which circumstances. This would not, however, constitute an explanation of the existence of that, or any, type of social institution. The presumption that the most efficient will tend to exist may be valid within an institutional framework of competition, property and contract, but cannot be assumed valid when that framework itself is in question. Similarly, the demonstration that all would benefit if all were to act in conformity to the rules of a social institution would not suffice to explain the existence of that institution. We cannot leap from collective benefits to collective action without violating the economic principle of methodological individualism. We must show that each would benefit if he were to act in the required manner.

This essay seeks to explain the nature and existence of social institutions in general, in terms of rational individual action and belief. I argue that such an explanation is possible, but is possible only if we shift our perspective from the rationality of actions per se to the rationality of rules of action. Social institutions arise in contexts where agents face the problem of the time-inconsistence of optimal plans under rational expectations. In such contexts it may benefit

individual agents to act as if they were committed to following a rule, the performance of an action specified by that rule being generally irrational except insofar as its performance is required to maintain others' belief that the rule is being followed. Social institutions are identical to agents' following, and believing others to be following, such rules.

If this thesis is correct, then economists who analyse action towards the institutional framework as if it were action within the framework are making a fundamental mistake. Within the framework the problem of time- inconsistency does not arise, and we can meaningfully speak of the rationality of an action per se. When the framework itself is in question it is meaningless to speak of the rationality of an action per se; one must instead consider the rationality of following a rule of action.

The distinction I am making between actions and rules of action is similar to the distinction between discretionary acts of policy and policy rules, as made by Kydland and Prescott (1977). It corresponds to my own previous distinction between teleological and deontological methods of formulating plans of action in Rowe (1982a). My refusal to speak of the rationality of an action without reference to the rule which requires it is based on motives similar to Lucas' (1980) refusal to evaluate a single discretionary policy action. We cannot evaluate an action without reference to its effect on expectations, and that effect will depend on the rules to which that action might, or might not, conform. What I seek to show in this essay is that the distinction between rules and discretion is not merely an abstruse part of the theory of macroeconomic policy, but is central to understanding any form of social life among rational agents.

The plan of the essay is as follows: for simplicity and concreteness we restrict our attention to one particular type of social institution - that of private property rights. It must be remembered, however, that our interest is not with property per se, but with property as an example of social institutions in general. With this proviso understood, we seek to answer two questions; what are property rights, and why do property rights exist?

The second section of this essay examines what should be regarded as constituting an answer to those two questions, and outlines the strategy adopted to provide the answers. The strategy adopted is not new. Like Hobbes (1651) and others, we posit an initial, pre-social "state of nature" and show how property rights might emerge from that state. What is new is our commitment to the economic approach and our explicit methodological analysis of the use of that strategy. Our "state of nature" is not intended as an equilibrium analysis of some actual historical state, but is posited as an imaginary disequilibrium state in the context of a stability experiment meant to represent, however abstractly, today's social world. The stability experiment is used to elucidate the forces which maintain our institutional framework in existence.

Section three examines what is involved in one agent's recognising another agent as a rational agent - such mutual recognition being a prerequisite for social interaction. It points towards the general problem of time-inconsistency of optimal plans under rational expectations.

Section four presents a specific model of the state of nature and reveals the inefficiencies of that state.

Section five argues that the mere revelation of those inefficiencies relative to a world of exogenously imposed property rights does not suffice as an explanation of property.

Section six notes that the state of nature differs from the actual world in that agents fail to use trust and deterrence to protect goods, and justifies our saying that property rights do not exist in the state of nature.

Section seven argues that the state of nature is not an equilibrium state, for agents therein fail to adopt rules of action in an environment where it would benefit them individually to do so. Eschewing discretion in favour of rules is required to establish trust and deterrence. By permitting agents to follow rules we attain an outcome different to the state of nature - the "social equilibrium".

Section eight justifies our saying that property rights exist in the social equilibrium, and hence completes our explanation of the nature and existence of property rights.

Section nine points out some remaining problems in the analysis and outlines ways in which the analysis could be extended.

The approach taken in this essay is similar to other work in the Contractarian tradition in that, following Hobbes' Leviathan, it posits an initial "state of nature" and analyses how social institutions might "emerge" from such a state. In comparing this to previous work in the same tradition, however, the reader should keep the following distinctions in mind. Firstly, we are conducting a positive analysis to explain why property rights in fact exist, in contrast to the primarily normative analysis of e.g., Hume (1777), Rawls (1971) and Nozick (1974). Secondly, like Hobbes, but unlike Locke (1690), there are no social institutions or rights of any kind in our initial "state of nature." We seek to analyse the "emergence" of individually enforced private property rights from a completely asocial state. Following the Lockean approach, Nozick, for instance, analyses the logically posterior question of the emergence of

collective enforcement of previously existing property rights. For another contrasting example, Demsetz (1967) analyses the emergence of private property rights from an initial position of common property rights while presupposing the institution of contract. Buchanan (1975) is perhaps the work closest in spirit to this essay, especially in its treatment of the inefficiencies of the state of nature and its recognition of the prisoner's dilemma aspect of that state. In my opinion, however, this essay constitutes an advance over that work in the following three major areas. Firstly, it provides an explicit methodological account of what the theorist is doing when he posits a "state of nature" and an "original contract" - he is performing a stability experiment. Secondly, it relates the problem of the original contract to the underlying problem of the time-inconsistency of optimal plans - the distinction between rules and discretion. Thirdly, it shows that the "original contract" can (and must) be enforced by the rules of action by rational individual agents. Rights can exist without collective enforcement, and collective enforcement cannot exist without individual enforcement of the collectivity.

II. Strategy

We seek to answer two questions. What are property rights? Why do property rights exist?

The first question is one of conceptual analysis. What does it mean to say that property rights exist? What sorts of action and belief on the part of a group of agents constitutes their having a system of property rights?

The second question is more a question of economic theory in the narrower sense. Why do individual agents perform those actions and hold those beliefs which constitute their having a system of property rights?

We will attempt to answer both questions within the context of a single conceptual experiment. With regard to the second question, our experiment can be seen as a stability experiment of a rather elementary type. The theorist who performs a stability experiment wishes to show that a certain variable is endogenous - being determined in equilibrium by other, exogenous, variables. To show this he posits an arbitrary shift in that variable holding the exogenous variables constant. He then seeks to show that, holding that endogenous variable at its new level by theorist's fiat, there would exist forces which, if permitted to operate, would tend to restore that variable towards its original level. The theorist who thus posits a state of affairs, in which the endogenous variable is at an "inappropriate" level vis-à-vis the exogenous variables, does not thereby assert that such as state of affairs ever existed at some historical time. He is explaining why that state of affairs does not, and could not, in fact exist.

It is this sort of experiment that Hume (1972) performs when he imagines an arbitrary shift in the (endogenous) money supply, and that

Patinkin (1965) performs when he imagines an arbitrary shift in the (endogenous) price level. We are doing exactly the same thing in this essay when we imagine an arbitrary non-existence of all social institutions such as property rights, the existence of which we seek to show to be the equilibrium level of some endogenous attribute of the world.

Following Hobbes (1651) we shall call the imaginary state of affairs in which no social institutions exist the "state of nature". But since our state of nature is no more than an imaginary state of affairs posited within the context of a stability experiment, it is inappropriate to ask whether the state of nature ever exists or existed at some historical time or place. The whole point of the exercise is to explain why the state of nature does not, and cannot exist. This is not to deny that at some times and places social institutions might not exist, any more than Hume or Patinkin would deny that the money supply or price-level respectively might be different at some times and places than what they are here and now. If the endogenous variable is different somewhere else, it is associated with different levels of the exogenous variables. Similarly, an actual state of affairs in which property rights do not exist cannot be the state of nature, for the latter, we seek to show, is a state of affairs in which the absence of property rights is <u>inappropriate or impossible</u> given the exogenous variables.

We need to examine the state of nature in some detail to examine the forces therein which would lead to the emergence of property rights and hence a dissolution of the state of nature. To do this, we place an <u>arbitrary constraint</u> on the behaviour of agents to prevent their taking the sorts of actions which would constitute their establishing property rights. We seek to show that they would wish to violate those arbitrary constraints. Our procedure is thus analogous to Barro and Grossman's

(1971) arbitrary fixing of prices at the "wrong" levels. They examine the "fixed price equilibrium" that would result if price adjustment were thus arbitrarily constrained, which enables examination of agents' desires to violate those constraints. Our own state of nature is analogous to their fixed price equilibrium. It is an equilibrium given those constraints, but since the imposition of those constraints is arbitrary we seek to show that it is not a genuine equilibrium - that rational agents would not in fact choose actions which fall within those constraints, but would wish to violate them. If we can show this - that individual agents would wish to violate the constraints that prevent their establishing property rights - then we have explained the existence of property rights.

Our strategy must be slightly more oblique than this however, for we seek also to explain the nature of property rights. If we do not know what actions and beliefs on the part of agents would constitute their having a system of property rights, we do not know what arbitrary constraints to impose on agents to preserve the state of nature as a constrained equilibrium. The constraint will be imposed implicitly, to be revealed later. We will simply posit a state of nature which resembles our intuitive perception of a state of affairs without property rights. We will then suggest alternative courses of action which rational agents would choose to take which, if taken, would lead to the establishing of a second outcome, the "social equilibrium", which resembles our intuitive conception of a world in which property rights do exist. In examining the constraints which would need to be imposed to preserve the state of nature and prevent the social equilibrium from being the outcome we will reveal the implicit constraints needed to derive the state of nature as

an equilibrium, and simultaneously discover the nature of the actions and beliefs which constitute the existence of a system of property rights.

The above paragraphs outline our basic strategy to answer the questions we have posed. Two futher points, however, remain to be emphasised concerning the nature of the task at hand.

Firstly, while it is not without interest to explain the existence of some particular social institution while taking as given the existence of other social institutions, that is not the question addressed in this essay. Demsetz (1967) for example, seeks to explain the emergence of private property from an initial state of common property while presupposing the institution of contract. Since this essay ultimately seeks to examine the compatibility of the existence of social institutions with the methodological assumptions of individualism and rationality, to populate the original position with agents who are anything but completely asocial would be to beg the question. We must imagine our agents as being plucked from so many desert islands where each has lived a previously solitary existence.

Secondly, while it is not without interest to explain the collective benefits that arise from the existence of property rights, such an explanation would not alone constitute an explanation of the existence of property rights in the required sense. We cannot introduce as deus ex machina some ill-defined entity which, by defining and enforcing property rights, solves a problem which rational individual agents cannot solve, and claim thereby to have answered our question. In terms of the methodology assumed here, if the problem cannot be solved by rational individual action then it cannot be solved at all. We cannot leap from collective benefits to collective action if we seek to examine the compatability of the existence of social institutions with the economic

approach to human behaviour. That all would benefit if all were to establish and respect property rights does not entail that each would benefit if he were to do so. Nor does it suffice simply to admit the possiblity of "transactions costs" in agreeing on and enforcing "collective action". The problem of agreeing on and enforcing "collective action" to agree on and enforce property rights is the same sort of problem as is the agreeing on and enforcing of property rights themselves. We would simply be shifting the problem from one place to another and using the term "transaction cost" as a name for our ignorance. We must show that the establishing of property rights is an outcome of rational individual action, that each would benefit if he were to protect and respect those rights.

III. Recognition of Intersubjectivity

The agents with whom we populate the initial position are previously solitary. They are solipsists, each having played a game against nature on his own lonely island. We will take these solitary agents and put them together on a "crowded" island, where their actions cannot but impinge on each other.

For any social institution to emerge it is necessary that each agent come to recognise the existence of other agents qua rational agents - as objects qualitatively distinct from rocks, plants, and other non-persons. How does this recognition come about?

The consequences of any plan of action, and hence the utility an agent can attain by implementing that plan, depend on the constraints facing him. Hence it will generally benefit an agent to invest resources in collecting information about those constraints so that he can formulate his optimal plan taking that information into account and thereby increase his (expected) utility. If the constraints facing an agent include the behaviour of another agent, it may thus benefit him to collect information in order better to predict and control that behaviour. The economic agent is in the same position as an economic theorist. He will need to construct a theory of his environment, which includes the behaviour of other agents.

If we assume that it is rational for us, as economists, to construct a theory of agents' behaviour by positing a set of preferences and beliefs and attributing rationality, then we may assume that the rational agent in the original position will do likewise, when he constructs a theory of his fellow rational agents. If he does this however, he must recognise that in a symmetrical informational environment the other agent

will construct a similar theory to predict and control his behaviour. To recognise another as a rational agent is to recognise that such recognition may be mutual. In this manner the rationality of both agents may become common knowledge between both agents.

Mutual recognition of rationality fundamentally alters the nature of the game played by the agents. They are no longer playing a game against "blind" nature, but instead a game against another rational agent whose moves depend on the theory he holds with respect to other players' likely moves. The utility that A can attain in this game depends upon B's moves, which depend on B's theory of A's moves. A can thus have an incentive to modify B's theory about A, and he also has an ability to modify A's theory, since, possessing free will, he can, if he so chooses, contradict by his actions any theory B might hold about his actions. What this means is that there exists no theory of A's actions which is true independently of A's wanting that theory to be seen as true.

We can thus distinguish two theories of an agent's actions. The first, or "basic" theory, correctly predicts an agents actions under the assumption that he takes as given others' theories about his actions. The second, "self-referential" theory, correctly predicts an agent's actions under the assumption that he rationally takes into account the effect of his actions on others' theories about his actions. I shall argue in this essay that the equilibrium that would result under the first theory of agents' actions is fundamentally asocial and corresponds to the Hobbesian State of Nature. The equilibrium that would result under the second theory is fundamentally social, and that social institutions are the outcome of agent's taking into account their ability to confirm or contradict others' theories of their behaviour.

IV. The State of Nature

If property rights are well defined, if enforcement and transactions costs are zero, and if free contracting (the exchange of rights) is allowed, then the Coase Theorm leads us to the conclusion that the equilibrium allocation must be efficient (Pareto Optimal). If the equilibrium allocation were not efficient, then agents could, and rational agents would, conduct additional mutually beneficial exchanges of rights. Seen from this perspective, the reason why the state of nature is inefficient is simply that there are no rights to exchange. Certain actions, viz the exchange of rights, are simply not options for agents in the state of nature.

This cannot be a final answer however. What does it mean to say that rights do not exist? What forms of physical behaviour and beliefs thereon are precluded to agents in the state of nature? How are those inefficiencies manifest?

To answer these questions we need a simple model fo the state of nature. We will take the model already provided by Skogh and Stuart (1978), adding extensions as we see fit.[2]

All agents are identical. Each has a fixed amount of a scarce resource, time, which he can allocate between three activities. The first activity is the production of goods. Time devoted to the second activity, "transfer", enables the agent to consume goods produced by other agents. Time devoted to the third activity, "protection", serves to reduce the quantity of goods transferred to other agents for a given amount of transfer activity on their part.

An agent's consumption of goods in any year is the sum of his production plus net transfers from other agents (we assume the good to be

perishable). An agent's utility in any year depends on his consumption in that year. He seeks to maximise the discounted sum of utilities, given a constant geometric rate of discount.

Having described the tastes and technology of agents in the original position, we will describe the state of nature which we put forward as a candidate for the equilibrium outcome of that world.

The state of nature is simply the solution that emerges from each agent's choosing his current allocation of time so as to maximise his current consumption taking as given the allocations chosen by other agents. In other words, the state of nature is simply the Nash Equilibrium.

If we assume sufficiently decreasing marginal returns to each activity then we will have an interior solution in which each agent allocates a positive amount of time to each activity and equalises the marginal return to each activity. Since all agents are identical, net transfers will be zero so that each agent's consumption of goods equals his production of goods.

This state of nature is inefficient in the following sense; an omniscient dictator could impose an allocation which could increase the utilities of all agents relative to the state of nature. A prohibition of transfer activity, together with the resultant elimination of individual agents' incentives to devote time to protection, would increase the amount of time devoted to production and hence increase production, consumption and utility.

Simple extensions to the above model can reveal additional sources of inefficiency. Let us introduce a second good, for instance leisure, which is easier to protect and harder to transfer than the first good, (call it "bananas"). The absence of a prohibition on transfer activity

reduces the private benefit to producing bananas relative to producing leisure, because the more bananas an agent produces the more will be transferred to others, or else the more time he will need to devote to their protection, and this will lead him to switch resources towards the good which is less vulnerable.

With the model thus extended, the efficiency losses in the state of nature are embarrassingly similar to the efficiency losses due to an "income" tax. Some resources are devoted to collecting the tax (transfer activity). Some resources are devoted to evading the tax (protection activity). Thirdly, the tax (if it is not a lump-sum tax) causes a misallocation of resources away from taxable activities (producing bananas) towards untaxed activities (leisure).

So far, all that is required to ensure an efficient allocation is the prohibition of transfer activity. The construction of an unclimbable fence around each agent's banana plantation could enable the dictator to achieve as efficient a result as would his choosing agents' allocations directly. This result is not general. If we further modify our model so that agents would have an incentive to exchange goods, were exchange possible, then the mere prohibition of transfer activity will not ensure efficiency, and may even lead to an outcome inferior to the state of nature. Let us therefore introduce a third good, "apples", and assume that agents have identical tastes, transfer and protection technologies but that type A agents have a comparative advantage in producing apples, while type B agents have a comparative advantage in producing bananas.

Obviously, the mere prohibition of transfer activity now no longer ensures efficiency of allocation, for this would condemn agents to self-sufficiency whereas, given our assumptions, efficiency requires that they specialise in production and produce a vector of goods different to the

vector they consume. Less obviously, the mere prohibition of transfer activity could possibly make agents worse off than in the state of nature. To see this, notice that type A agents will devote more time to the transfer and protection of bananas than will type B agents, because the alternative source of consuming bananas - production - is less rewarding for type A than for type B agents. The reverse is true for apples. Therefore, in the state of nature there will be net transfers of bananas from type B to type A agents, and net transfers of apples from type A to type B agents.

Under the strong assumptions that consumption of both goods is necessary for life, and that type A agents can produce no bananas, and type B agents can produce no apples, it immediately follows that agents are better off in the state of nature than they would be under a prohibition of transfer activity.

In the absence of voluntary exchange, forced exchange may be a Pareto Improving activity, for agents will tend to "steal" the goods they value most. Forced exchange is inefficient relative to voluntary exchange insofar as, in the former case, resources are devoted both to effect and to prevent transfer, so that more resources are used to yield a given amount of net transfer.

With our model of the original position thus extended, our benevolent dictator can no longer ensure an efficient allocation merely by building an unclimbable fence around each agent's plantation or orchard. If he insists on meddling with the technology of the original position, rather than simply choosing an allocation by dictatorial fiat, he would need to build in to each fence a clever little device called an "exchange mechanism". An exchange mechanism is a pair of boxes, one each side of the fence. Goods can be placed in each box, and if and only if agents

pull a lever simultaneously, will the boxes both swivel to the opposite sides of the fence.

V. Prisoners' Dilemna

The preceding section describes the state of nature as the Nash Equilibrium of the original position. We have seen how a benevolent dictator, either directly by commanding an allocation, or indirectly by constructing fences and exchange mechanisms, could effect an improvement of agents' welfare compared to the state of nature. To ensure an efficient allocation directly by command, the dictator would, in general, need to know the exact numerical values of the parameters of the functions describing agents tastes and technologies. If the dictator opts for the indirect method of constructing fences and exchange mechanisms, the requirements on his omniscience are far less strict. He need only know what variables enter as arguments in the taste and technology functions.[3]

The original position, once the fences and exchange mechanisms are in place, provides an exact mechanical counterpart to the world presupposed by the Coase Theorem. Property rights (the fences) are exogenously defined and inviolable. Transactions costs (if any) can be represented by assuming it requires effort to pull the levers on the exchange mechanisms. Any misallocation of rights, (any misallocation of goods between the compounds) will send rational agents running to exchange rights (running to the exchange mechanisms) to achieve a new, Pareto Optimal, allocation.

If the allocation in the state of nature corresponds to the allocation in a world without property rights, and if the allocation under fences and exchange mechanisms corresponds to the allocation in a world of private (alienable) property rights, then we have revealed the inefficiency of the allocation in a world without property rights relative to a

world with property rights. In doing this, however, we have not explained either the existence or the nature of property rights.

The dictator, with his commands or fences and exchange mechanisms, is an obvious deus ex machina brought in by theorist's fiat to save the players from their predicament in the state of nature. That all would benefit if all were to act as if the dictator existed does not entail that each would benefit if he were to act as if the dictator existed. Taking others' actions as given, as is done in deriving the Nash Equilibrium for the state of nature, it is obvious that each does not benefit by acting as if the dictator existed. Moreover, in building fences and exchange mechanisms, the dictator does not thereby create property rights; he merely changes the technological constraints facing agents in the original position in such a way that property rights become unnecessary.

The state of nature is inefficient in this sense; each agent would like to exchange his commitment to act as if the fences and exchange mechanisms existed in return for a similar commitment by other agents. If such an exchange were feasible, it would be undertaken and all agents would become better off. The problem is; is that exchange feasible? What reason could each agent have for believing that others will keep their side of the "original contract" if and only if he keeps his side of the bargain? Taking as given the others' decisions whether to act as if fences and exchange mechanisms exist, each agent's best reply is always to act as if they do not exist, for they do not in fact exist.

The state of nature thus corresponds to the "noncooperative" solution to the game of prisoners' dilemma, a statement which is not surprising given that the latter game is the prime example of the purported

inability of rational individuals to attain a mutually beneficial outcome.

VI. The Asocial State

If our stability experiment is to be successful in explaining the existence of property rights, we must show how property rights emerge in equilibrium from rational individual behaviour in the original position. If it is accepted that property rights do not exist in the above-described state of nature, then if our experiment is to be successful we must show that the state of nature is not in fact an equilibrium outcome. Somehow we must have slipped in an implicit assumption which prevents agents from acting in a way they would wish to, and which would lead to the emergence of property rights. What is this implicit assumption? How could an individual agent increase his utility by departing from the strategy he adopts in the state of nature?

To answer this question and to understand the distinction between a state of affairs with and without property rights, we should compare the way that agents protect goods in the state of nature with the way they protect goods in the actual world.

Agents in the state of nature protect goods by actions which take place in advance of, or at the time of, the transfer activity they seek thus to counter. Their ability to protect goods in this manner is a mere technological datum. It depends not on ownership of the goods protected, but on a merely physical relation between the agent and goods.

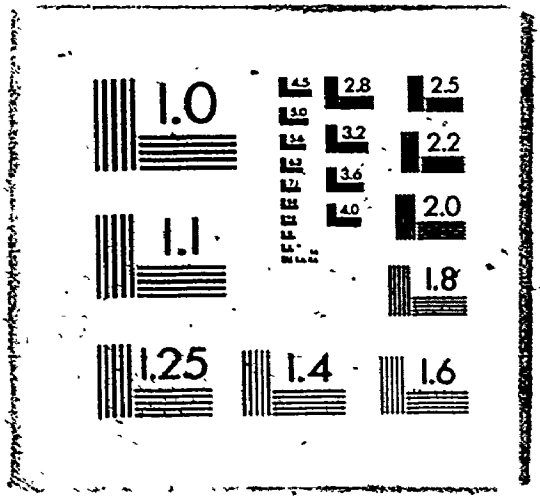
We see agents in the actual world use similar sorts of actions to protect goods from consumption by others. Concealment, and the installation of fences, locks, and armed guards are examples of such actions the effectiveness of which is a purely technological matter. But this is not the only way in which agents in the actual world protect their property. They not only take prior action to forestall the possibility of future

theft, but also promote expectations of action to follow after, and contingent upon, any theft which might have taken place. They use deterrence. The ex ante expectation of the ex post act act of punishment can prevent theft even though the ex post act itself cannot. Deterrence is the primary method of protection in the actual world. Other devices, such as fences and locks, are used only to the extent that identification of the criminal, and hence punishment, is prohibitively costly.

Natural predators, which do not have expectations, cannot be deterred. Rational predators, which do, can be deterred. In using deterrence as a method to protect his goods from others, an agent ipso facto recognises that he is playing a game, not against nature, but against other rational agents.

This is what provides our justification for saying that property rights do not exist in the state of nature. A collection of agents, each of whom perceives himself as a solitary rational agent, cannot be said to have a system of property rights. Agents in the state of nature, in failing to use deterrence, protect goods from other agents in fundamentally the same way as they would protect goods from the ravages of nature. Acting thus, any ascription of rationality to other agents - any recognition of other rational agents as rational agents - would be redundant. In other words, they act as if they believe themselves to be solitary, and as such, cannot be said to recognise property rights nor indeed to recognise any social institution.

2



VII. Escaping the State of Nature

Property rights do not exist in the state of nature. If we are to explain the existence of property rights we must show that the state of nature is not, in fact, an equilibrium - that agents therein are not acting rationally. What mistakes are they making? The mistake made by each agent in the state of nature is that he lets bygones by bygones. In doing so, he curtails his ability to influence to his advantage the expectations, and hence actions; of other agents.

Since the constraints facing an agent depend on the actions of other agents, those constraints are not independent of the expectations of other agents concerning his own future actions. Alternatively, using the more familiar terminology of Lucas (1976), each agent is like the policy-maker who faces an economic structure which is not invariant to other agents' beliefs about the policy rule being followed. Agents in the state of nature fail to recognise and exploit that fact.

If the agent in the original position could choose conspicuously to commit himself to any technically feasible policy rule, the rule which he would rationally choose would specify a historical strategy - i.e., the action specified at each point in time by the optimal rule would be contingent, in an irreducible manner, on past events. As a result, an action specified by the optimal rule could well be irrational at the time of its performance, except insofar as its performance is necessary to maintain others' belief that the same rule will be followed in future.

We have already established that the state of nature is inefficient insofar as each agent would like to exchange his commitment to act as if the dictator's fences and exchange mechanisms existed for a similar commitment on behalf of other agents, were such an exchange feasible. It is

the perceived commitment to historical strategies which can make that exchange feasible. By being believed to have adopted a historical strategy which rewards past adherence to, and punishes past violations of the "original contract", an agent can provide other agents with sufficient incentive to adhere to the original contract, at least where he himself is concerned.

In deriving the psuedo-equilibrium state of nature, we supposed that, since goods are perishable and hence one year's exogenous physical facts are independent of last years' outcome, we can analyse the original position as if it lasted but a single year, having no past or future. That supposition is illegitimate insofar as it eliminates, without argument, the possibility that agents might rationally adopt historical strategies and be perceived to do so.

The assumed tastes and technologies of the original position already contain the possibilities for agents to punish and reward the actions of other agents. Agent A can reward agent B by refraining from engaging in unilateral transfer activity directed at the goods B has produced. A can punish B by engaging in "large" amounts of unilateral transfer activity directed at the goods B has produced. In order to follow a historical strategy which punishes or rewards B contingent upon B's past transfer activity towards A, all that is required is that A be able to monitor B's past transfer activity towards himself. We will assume that requirement satisfied.

Faced with A's historical strategy, when considering the amount of transfer activity he will direct towards A's goods, B must consider not only the effect on his current consumption (as he does in the state of nature), but also the discounted loss of future consumption caused by his triggering future punishment from A. Provided that B's rate of discount

is not too high, by setting the punishment large enough, in intensity or duration, A can ensure that the discounted value to B of engaging in any amount of unilateral transfer activity directed at A must be negative. If A's strategy is believed by B, then A can completely deter B from violating his "original contract" with A.

The state of nature is inefficient insofar as each agent would like to exchange his commitment to act as if fences and exchange mechanisms existed, for a similar commitment by other agents. By following and being believed to follow a historical strategy of punishment and reward, an individual agent can ensure that he and another agent will act as if this exchange of commitments had been made. In other words, an individual agent can police the "original contract" between himself and other agents by following and being believed to follow a historical strategy of the required type.

This conclusion does not depend on the assumption that the number of agents is small. On the contrary, the incentive an agent has for being believed to follow a historical strategy of punishment may actually be higher in a large numbers game, since only a small increase in punishment will be required to persuade many contract violaters to switch to raiding other agent's plantations. What is required is that an agent be able to identify which agents have, and have not, violated their "original contracts" with him. If identification were not possible, any punishment would have to be indiscriminate, and would fail to deter if each agent perceived his being punished as independent of his own decision to violate, as would likely be the case in a large numbers game.

An agent who is believed to be following a historical strategy protects his goods by trust and deterrence. Others trust that if they do not violate the "original contract", then neither will he. They are

deterred by the belief that if they do violate the contract, then he will punish them in subsequent games. They believe that their current choice of action will influence his future action, even though they know that the underlying "physical facts" - the basic exogenous parameters of tastes, technology, and information thereon - are in no way influenced by the outcomes of previous years' games. Is this belief rational? Would it in fact be rational, at each point in time, for an agent to adhere to a historical strategy, letting his current choice of action depend on the "irrelevant bygone" of last year's outcome? Would it be rational e.g., for A to play "tit for tat", taking bananas from B this period if and only if B took apples from A last period?

Let us understand the hypothesis of rational expectations to mean that B's expectation of A's action is identical to the action that A will rationally plan to perform. In year one, A will plan his action in year two taking into account the effect of his plan on B's current expectation and hence on B's current action. A will rationally choose a plan which makes his future action contingent upon B's current action. When year two arrives, however, B's action in year one is predetermined, and hence cannot be influenced by A's action in year two. A may therefore reformulate his optimal plan of action in year two, this time taking B's action in year one as given, as indeed it is. If A permits himself to reformulate his optimal plans in this way, then his plans will be time-inconsistent. The (contingent) action he plans in year one to perform in year two will not coincide with the action he plans in year two to perform in year two. If B knows that A adopts this method of choosing plans of action, he will know that only the present portion of A's plans will ever be carried out - he will not rationally believe the implied promise of A's currently planned future contingent actions, since he knows that

the only incentive A has to promise is the effect his doing so might have on B's current action, an incentive which disappears at the time the promised action is to be performed, for B's action is past and cannot be influenced. Unable to influence B's expectations, A's motive for adhering to a historical strategy collapses, and the outcome reverts to the state of nature.

This method of formulating plans of action, whereby the choice of current plan is discretionary, and need not conform to the agents' previous optimal plans, I have elsewhere termed the "teleological method", in Rowe (1982a). An agent who adopts the teleological method will not follow a historical strategy, for it is only his past optimal plans which require his current action to be irreducibly contingent upon then - future, but now - past events, and past optimal plans are irrelevant to a teleological agent. The state of nature is thus the outcome when agents adopt the teleological method. The fundamental reason why life in the state of nature is "solitary, poor, nasty, brutish, and short" is that agents therein adopt the teleological method in an environment in which their optimal plans under rational expectations are time-inconsistent.

To escape from the state of nature, each agent would like conspicuously to commit himself to a historical strategy, to promote the expectations required for trust and deterrence. The original position, however, contains no technology which enables an agent thus to commit himself. He lacks Ulysses' bonds. Would it nevertheless be rational for him to act as if he were thus bound, in order to promote the expectation that he will maintain that same stategy in future? Could it ever be rational for B to believe that A will act <u>as if</u> he were constrained to follow his past optimal plans?

It is these two related questions which I sought to answer in Rowe (1982a). To adopt a historical strategy an agent must eschew discretionary action in favour of following a rule; he must act as if he were committed to fulfill his past optimal plans - he must adopt the "deontological" method of formulating plans of action. It is rational for an agent to adopt the deontological method only if his doing so, or failing to do so, can influence the expectations of other concerning the likelihood of his doing so in future. It is rational, in other words, to be bound by the implicit promise of one's past optimal plans only to the extent that doing so is necessary to promote the expectation that one will keep promises in future. If expectations of future actions depend only on the basic parameters of tastes, technology and information thereon, then expectations cannot be influenced by past actions, for we have assumed each year's basic parameters to be invariant to previous years' outcomes. In adopting the historical strategies implied by the deontological method, however, an agent's actions violate the correspondence between current actions and basic parameters and hence destroy the possibilty of other agents forming their expectations in this manner. Other agents are then forced to look to past actions as a guide to future actions, and can only interpret the actions of an agent who adopts the deontological method on the surmise that this is exactly what he intends them to do, for if he succeeds in this intention then it is indeed rational for him to follow the deontological method, and hence rational to expect him to do so.

By following, and being believed to follow, a historical strategy of punishment and reward, an individual agent can ensure an outcome which is as if he had exchanged a commitment to act as if the dictator's fences and exchange mechanisms existed, for a similar commitment towards him on

the part of other agents. The rationality of following a historical strategy depends on the rationality of adopting the deontological method - of adhering to the promises implied by one's past optimal plans. The keeping of promises can promote the expectation that one will do so in future, as I have argued in the above paragraph and, in more detail, in Rowe (1982a). It is rational to keep one's promises insofar as the value of one's reputation (the present value to the agent of the expectations thus promoted) exceeds the benefits of a current violation.

We have thus demonstrated the possibility of an equilibrium in which rational individual agents act as if the dictator had constructed fences and exchange mechanisms. What maintains this "social equilibrium", however, is not a fictional dictator but the rational adoption, and rationally perceived adoption, of rules of action which require agents' obedience to their past optimal plans.

VIII. The Social Equilibrium

In the social equilibrium, all agents act as if our imaginary dictator had constructed an unscalable fence around each agent's plantation or orchard, and had installed exchange mechanisms within that fence. They engage in no protection activity as such - the whole of each agent's produce being apparently free for the taking by other agents. Agents engage in transfer activity, taking from others only the goods he himself does not produce, whether apples or bananas. Such transfer activity requires but small cost of time, for it does not have to overcome protection activity. Agents' transfer activities are peculiarly limited however, given the apparent ease with which each could increase his consumption by increasing the amount of transfer activity he directs at both goods.

Agents' actions, in other words, seem to bear no relation to the physical facts of their environment. For example, an agent will consume the bananas that he has produced, but will avoid consuming the identical bananas that his neighbour has produced, despite the lack of physical barriers to his taking the latter.[4]

It is this divorce of actions from the physical facts of their natural environment which justifies our saying that agents in the social equilibrium recognise a social institution - that property rights exist in the social equilibrium. Each recognises the bygone physical act of past production as constituting the social fact of the relevant agent's current ownership of those goods, until a mere pair of physical gestures, the pulling of the levers on the imaginary exchange mechanism, constitutes the social fact of the exchange of property rights.

If we should ask why agents in society act with regard to more than the merely physical constraints of their environment, we must refer to

the beliefs that each holds regarding the likely effects of a contrary course of action upon the expectations and future actions of his fellows, and these actions and expectations hinge upon the rules of action which agents follow and are believed to follow. An agent will not violate the rule of property insofar as he fears the effect his doing so would have upon others' beliefs concerning the rule of action he is following, and fears the effect of his doing so upon the future actions of others given the rules of action - the historical strategies - he believes them to be following. In saying that an agent does not steal because he fears retribution and the loss of trust we make implicit reference to the practice of deontological methods of action - of obeying past optimal plans.

Social institutions, and the social facts to which they give rise, are none other than theoretical constructs, shared by social theorist and social agent alike, which, like our imaginary dictator's fences and exchange mechanisms, permit us to make as if explanations of the actions and beliefs of agents in contexts where the following of rules can be taken for granted by both agents and theorist. The anticipated penalties for violation of the "rules of the game" which define action within a given framework of social institutions, and are taken as exogenous to the game within that framework, are themselves endogenous to the wider pre-social noncooperative game, and are created by the rules of action adopted by rational individual agents playing that wider game.

IX. Conclusion

The perspective of economic theory is a forward-looking or teleological perspective. Rational agents act the way they do in order to attain a desired end. In contrast, many of the institutions which define economic activity have an inherently backward-looking or deontological perspective. Payment is owed because of debts undertaken in the past. Criminals are punished because of past crime. Property is owned because of past acquisition. An agent who looks at only present and future simply cannot see such social (and moral) facts, and yet such facts are seen, for without their generally being seen they would not exist. How can we reconcile the forward-looking perspective of economic theory with the existence of social institutions which create such backward-looking economic facts?

The clue to this reconciliation can be found in a paradox of teleological rationality - the problem of the time-inconsistency of optimal plans under rational expectations. An agent's current optimal plan for future anticipated actions does not generally coincide with his future optimal plan when those anticipations are predetermined. In such contexts it may be rational, on teleological grounds, for an agent to renounce teleological rationality in favour of deontological rationality - acting in conformity to the promises implied by his past optimal plans.

Just such a reconciliation has been attempted in this essay, in explaining the nature and existence of property rights as an instance of social institutions in general. The approach taken is basically contractarian, following in particular Hobbes' Leviathan, in positing a "state of nature" and an "original contract". Care has been taken, however, to

elucidate the methodological role of such constructs, and to examine the self-enforcing properties of the "original contract". The state of nature is posited as part of a stability experiment, not as a historical state. The "original contract" provides only an "as if" description of the process whereby the adoption of rules of action by rational individual agents prevents the state of nature from being actualised. The state of nature exists only as an ever-present logical possibility - a state of afairs the existence of which is posited only that its non-existence might be explained. The original contract was never made, but it is as if it were now maintained.

I wish now to point out three problems in the analysis which deserve closer study than has been given in this essay. The first concerns the uniqueness of the social equilibrium. The second concerns the stability of the social equilibrium given the rules followed therein. The third concerns the possibility that trust, reputation, and hence the enforcement of property rights, might be less than perfect in the social equlibrium.

I have argued that, compared to the state of nature, each agent would be better off under the "original contract" which required each agent to act as if an imaginary dictator had built fences and exchange mechanisms around each agent's plantation. There may, however, be several "original contracts", each Pareto Superior to the state of nature, and each Pareto Optimal. Alternative contracts, for example, might require some agents to pay "tribute" to others. Perhaps many such original contracts could be self-enforcing, in which case the social equilibrum is not unique. I suspect that this problem is not genuine, but that, like many "problems" with contractarian theories of society, it is due to mistaking a stability experiment for historical analysis.

Which particular allocation of property rights exists at a point in time is a question of history - it is an outcome of a historical process of production, exchange, settlement, (and indeed theft), the origins of which are lost in time. The question addressed by this essay is not which particular allocation will exist now, but how any current allocation of property rights, or property rights themselves, can be maintained. Jumping from one social equilibrium to another in real time is simply not an option. Attempting to do so would merely be to violate the very rules of action which maintain any social equilibrium. The expectations needed to sustain those new rules could not be present if the new rules were a violation of the old. A "society" which permitted purely discretionary redistribution of property rights would recognise no property rights at all, if indeed it could appropriately be called "a society" at all.

There are two senses in which the term "social equilibrium" might be used. The first sense refers to the set of rules of action, historical strategies or reaction functions that agents adopt. The second sense refers to the outcome, or actions of agents, given that those rules are being followed. It is the stability of the social equilibrium in the second sense that is our second problem alluded to above.

Suppose that all agents are following the rule of adhering to the "original contract" and punishing violators. If we start from a year in which no agent violates, then the rules provide the threat of punishment sufficient to deter future violation. Consider what would happen, however, if agent A mistakenly believes that agent B has violated the contract. In order to maintain the credibility of his rule, A must punish B for his violation, but in punishing B, A himself may be perceived by B as violating his side of the contract. B in turn may then

punish A, and the outcome degenerates into a mutually destructive vendetta, possibly far worse for both sides than even the state of nature. In order to ensure against the possibility of such an occurrence, agents must either ensure that they have certain knowledge that an agent they punish did indeed violate, or else find some way of signalling that their act of punishing is motivated solely by retribution, and is not motivated solely by the prospects for discretionary gain at the expense of the punished. Since certain knowledge is not readily available, agents must ensure that any act of punishemnt is costly not only to the punished, but is seen to be costly to the punisher. The implications of this problem are explored in Rowe (1982d).

The emphasis of this essay has been on demonstrating the possibility of trust, reputation, and hence of property rights. As a result of this emphasis, the possibilty that there might exist conditions under which trust, reputation, and property rights are less than perfectly sustained in equilibrium, has been given less than due attention. I make no apologies for postponing consideration of this, our third problem mentioned above. The concept of "theft", as a category of action, is meaningful only against a massive presuppositional background wherein theft is the exception. There can be no such thing as "theft" in the state of nature.

If a rule is to deter theft, it must set the expected cost of punishment to the potential thief greater than his expected benefits from stealing. If an agent is to enforce such a rule, it must be the case that the value to him of deterring theft, less the costs of monitoring, exceeds the costs of punishing. In this essay we have tacitly assumed the conditions necessary for some such rule to exist-low monitoring costs, low discount rates, long expected time horizons, options for

imposing high penalties at low cost, etc. If those conditions are not met, then the original contract will not be enforced. In Rowe (1982c) we examine the implications of this problem of the enforceability of contracts on prices and quantities traded in exchange contracts.

Footnotes

* I would like to thank Rosslyn Emmerson, Joel Fried, David Laidler, Tim Lane, Michael Parkin, Tom Rymes, and Ron Wintrobe for their valuable comments and suggestions on earlier versions of this paper. The responsibility for remaining errors and muddles is my own.

1. By comparing property rights with superstitions, David Hume succeeds admirably in conveying the essential "strangeness" of our observing the former:

"A fowl on Thursday is lawful food; on Friday abominable; eggs in this house and in this diocese are permitted during Lent; a hundred paces farther, to eat them is a damnable sin. This earth or building yesterday was profane; today, by the muttering of certain words, it has become holy and sacred....I may lawfully nourish myself from this tree; but the fruit of another of the same species, ten paces off, it is criminal for me to touch. Had I worn this apparel an hour ago, I had merited the severest punishment; but a man, by pronouncing a few magical syllables, has now rendered it fit for my use and service...But there is this material difference between superstition and justice, that the former is frivolous, useless and burdensome; the latter is absolutely requisite to the well being of mankind and existence of society." Hume (1777, pp. 197-198).

2. An alternative model of the state of nature is provided by Bush (1972).

3. See Hayek (1945).

4. See footnote 1.

References

Barro, R.J., and H. Grossman, "A General Disequilibrium Model of Income and Employment", American Economic Review, May 1971.

Berger, P.L. and T. Luckmann, The Social Contruction of Reality: A Treatise in the Sociology of Knowledge. New York, 1966.

Buchanan, J.M. (1975), The Limits of Liberty, University of Chicago Press: Chicago.

Bush, W.C. (1972), "Individual Welfare in Anarchy" in G. Tullock (ed.), Explorations in the Theory of Anarchy, Center for the Study of Public Choice; V.P.I.

Demsetz, H., "Towards a Theory of Property Rights", American Economic Review, 1967.

Hayek, F.A. (1945), "The Use of Knowledge in Society," American Economic Review, vol. 35.

Hobbes, Thomas (1651), Leviathan, Bobbs-Merril, 1958.

Hume, David (1742), "Of the Balance of Trade", Essays Moral, Political and Literary, London: Oxford University Press, 1963.

Hume, David (1777), An Enquiry Concerning the Principles of Morals, reprinted in H.D. Aiken (ed.), Hume's Moral and Political Philosophy, Hafuer; New York, 1968.

Kydland, F.E. and E.C. Prescott, "Rules Rather than Discretion: The Inconsistency of Optimal Plans", Journal of Political Economy, 1977.

Locke, John (1690), An Essay Concerning the True Original Extent and End of Civil Government in E. Barker, ed., Social Contract, Oxford University Press; London, 1969.

Lucas, R.E. Jr. (1976), "Econometric Policy Evaluation, A Critique", in Brunner and Meltzer, (eds.) The Phillips Curve and the Labour Market. Amsterdam: North Holland.

Lucas, R.E. Jr. (1980), "Rules, Discretion, and the Role of the Economic Advisor" in S. Fischer ed., Rational Expectations and Economic Policy, University of Chicago Press.

Nozick, R. (1974), Anarchy, State and Utopia, New York: Basic Books.

Patinkin, D., Money, Interest and Prices, 2nd Ed., Harper and Row, 1965.

Rawls, J. (1971), A Theory of Justice, Cambridge: Harvard University Press.

Rowe, P.N. (1982a), "The Rationality of Promises", mimeo.

Rowe, P.N. (1982c), "Trust, Wages, and Employment", mimeo.

Rowe, P.N. (1982d), "A Theory of Strikes", mimeo.

Skogh, G. and C. Stuart (1982), "A Contractarian Theory of Property Rights and Crime," Scandinavian Journal of Economics, vol. 84.

Chapter Four

Trust, Wages, and Employment*

I. Introduction

Why, given an apparent excess supply of labour, would profit- maximizing employers not reduce wages? One possible answer to this seemingly rhetorical question has been proposed by Calvo (1979), Eaton and White (1982), and Gintis and Bowles (1981); employers may rationally choose to set wages above market clearing levels in order to reduce worker malfeasance. The higher the wage the more the worker will lose if he is caught cheating and fired, and so the more likely is his cheating to be deterred. The profit-maximizing wage may thus exceed the workers's opportunity cost of entering the job, thus creating an equilibrium excess supply of labour.

This paper provides an integrated and more general treatment of the theory proposed in the above-mentioned papers. We first discuss why methods other than high wages might not be used to prevent cheating. Next, assuming that other methods are not adopted, we analyse the worker's choice-problem (whether or not to cheat), and the firm's choice-problem (whether to deter cheating, what levels of wages and monitoring to set), to provide an integrated framework for subsequent discussion. We then examine the circumstances under which the deterrence of cheating results in unemployment as opposed to non-compensating wage-differentials between identical workers employed in different firms. The explicit treatment of worker's and firm's decision-problems, with the level of monitoring endogenous, allows us to derive unambiguous comparative-static results showing how the (real) wage and (natural) rate of unemployment are affected by changes in the exogenous variables.

The use of a framework with an explicit time-structure, with workers both entering and leaving employment, enables us to show that lower turn-

over reduces the firm's costs of deterring cheating, thereby motivating long-term relationships between firm and worker, and also explains the practices of letters of reference and of employing bondable workers as substitutes for reducing turnover. Introducing worker turnover, and the insight thereby gained into labour market organization, enables us to relate the analysis of malfeasance in labour markets to Klein and Leffler's (1981) analysis of malfeasance in product markets. If the high wages (prices) needed to deter cheating attract the successful entry of new workers (firms) into the market, it is the resulting higher turnover of workers (firms) that can destroy equilibrium. If the organizational structure of the market can preserve low turnover despite the attraction of high wages (profits), then the sunk-costs invoked by Klein and Leffler are not needed to preserve equilibrium. Introducing turnover also allows us to disprove Eaton and White's assertion that the excess supply of labour which results when firms deter cheating allows employers to discriminate on economically-irrelevant characteristics at zero cost.

The final section of the paper shows how the results of previous sections would need to be modified if the limited use of other methods to deter cheating is possible.

II. Enforcing Exchange

All exchanges present opportunities and incentives for one or both parties to deliver less than he had led the other to expect. Each gives up something that he values, only because he expects that his doing so is a necessary and sufficient condition for getting the other to give up something that the other values. Making the Nash assumption that the other's action is given, each agent is obviously better off if he does not give up something he values. A mutually beneficial exchange can thus fail to take place for exactly the same reason that the prisoners in prisoners' dilemma fail to reach the mutually preferred (cooperative) solution. Rather than wonder why some apparently mutually beneficial exchanges do not take place (e.g. why unemployment exists), we should first be surprised that any exchanges are ever made.

Rational agents, realizing that a mutually beneficial exchange cannot otherwise take place, will seek to destroy the Nash assumption - to convince each other of the biconditional relation between giving and receiving (you get if and only if you give). To do this, agents must demonstrably alter the incentives they face, each to convince the other that he will maximise utility by giving if and only if the other gives. Clever mechanical devices might accomplish this task (each agent attaches an explosive device to his head which conspicuously will detonate automatically if he receives without giving) but these will presumably be costly to construct. Alternatively, they might seek the intermediation of a third party (representing "the state") who will enforce the contract by being known to inflict sufficiently large penalties on a defaulting first or second party.

Third parties, like mechanisms, are not costless to use. The third party must demonstrate that he has sufficient power to inflict a sufficient penalty on a defaulter, and is able to observe whether default has occurred. Moreover, third parties, unlike mechanisms, have interests of their own. Why should the third party, his own fee safety pocketed, bother to punish the defaulting party? Might he not rather share the illicit gains of the defaulter? The third party has made an exchange contract with the first and second parties to enforce their exchange, but exactly the same sort of problems pertain to the former exchange as to the latter. To invoke intermediation by third parties is not to answer but merely to postpone answering how exchange is enforced. To say that property rights make exchange possible does not suffice, for what is a system of property rights but an exchange whereby each refrains from performing certain actions, expecting for some reason that his refraining is a necessary and sufficient condition for others' refraining.

If we rule out mechanisms on the grounds of cost, and adopt a methodological prohibition against invoking third parties, to avoid circular reasoning, how are the first and second parties to enforce their exchange?

Stigler and Becker (1974) propose that an employer who seeks to deter worker malfeasance should require the worker to post a bond - the bond being forfeited if he detects the worker cheating. This is no solution, however, if the worker cannot trust his employer not to confiscate the bond arbitrarily. Posting a bond does not eliminate the need for trust, if simply shifts the need for trust from one party to the other. In place of a bond, could not the worker perhaps yield some hostage which is of no value to the employer? Eaton and White argue that collateral constraints may prevent workers from posting bonds and, by implication,

from yielding hostages. They have no bonds or hostages to give. Apart from collateral constraints, workers may be unwilling to yield hostages because doing so could make them vulnerable to blackmail. Unscrupulous operators could pose as employers, collect hostages, then reveal their true identity and demand payment for the return of the hostages. If both sides yield hostages, then the exchange of hostages itself needs enforcing.

These above considerations motivate our assumption that workers cannot yield hostages or post bonds, until the final section of this paper where we partially relax that assumption. If we rule out mechanisms, third parties, bonds and hostages, the only remaining incentive an agent may have for honouring his exchange commitments is that if his failure to do so is observed he may lose his trading partner's trust and thus lose some future exchange opportunities. An agent will cheat if and only if his expected loss of the surplus from future exchanges falls short of the immediate gains from cheating. What distinguishes enforcement by reputation from other forms of enforcement, therefore, is a direct link from the prices and quantities exchanged (and hence an agent's surplus or gains from future exchanges) to the ability of agents to enforce exchanges. Some exchanges, where the surplus from future exchanges cannot suffice to enforce exchange, will not be made. Others will be made, but at prices and quantities being set with a view towards enforcing exchange rather than clearing markets.

III. Wage Differentials

Let us assume that there are two types of firm (or job) in the economy: those that present opportunities for "cheating" and those that do not. For reasons which will be apparent later, I refer to the former as "good" jobs, and the latter as "bad" jobs. All jobs within each type are identical. The labour market is atomistically competitive, and all agents have perfect information on all relevant variables, except that it is costly for firms to monitor a worker's cheating. We assume that the only penalty a firm can inflict on a cheating worker is to fire that worker. All workers are identical, risk neutral, and have zero rates of time preference.

The Worker's Decision

Consider a worker currently in a good job. He has an opportunity to cheat his employer, and must decide whether to exploit that opportunity. He will follow that strategy which maximizes his average expected income.

If he decides not to cheat, then he receives a wage w per period while in the good job, and faces a probability d per period of dismissal. If he decides to cheat, he gains θ per period from cheating, plus his wage w, and faces a probability p per period of being caught cheating and fired, and a probability d per period of being dismissed for other reasons. A fired or dismissed worker can accept immediate employment in a bad job at wage v, and faces a probability s per period of regaining a good job.

A non-cheating worker expects to earn w for 1/d periods then v for 1/s periods. His expected average income per period is thus:

$$NC = \frac{ws + vd}{s + d} \qquad (1)$$

A cheating worker expects to receive $w + \theta$ for $1/(d + p)$ periods, then v for $1/s$ periods. His expected average income per period is thus[1]:

$$C = \frac{(w + \theta)s + v(d + p)}{s + d + p} \qquad (2)$$

The worker will not cheat if and only if $NC > C$. Firms can therefore deter workers from cheating by setting a wage, w^*, such that:

$$w^* > v + \frac{\theta}{p}(s + d) \qquad (3)$$

The Firm's Decision

The individual firm must decide whether or not to deter cheating, and if so, what level of monitoring, p, to choose.

Let us assume that the costs per period of monitoring, M, are proportional to the frequency of monitoring:

$$M = mp$$

If the firm decides to deter cheating, it must choose w and p to minimize the costs of wages plus monitoring per period.

$$M + w^* = mp + v + \frac{\theta}{p}(s + d) \qquad (5)$$

$$\frac{d(M + w^*)}{dp} = \frac{m - \theta(s + d)}{p^2} = 0 \qquad (6)$$

If cheating is deterred, the optimal level of monitoring is thus:

$$p^* = (\theta(s + d)/m)^{\frac{1}{2}} \qquad (7)$$

If the firm decides not to deter but to permit "cheating," then it need pay a wage, w, that is barely sufficient to attract workers away from bad jobs. Since the gains from "cheating," θ, are now a legitimate perquisite that goes with the job, it need only pay:

$$w = v - \theta \qquad (8)$$

Let the costs to the firm of a worker's "cheating" be c per period. The firm will minimise its total costs per period per worker by deterring cheating if and only if

$$w^* + M < v - \theta + c \qquad (9)$$

Substituting from equations 3, 4, and 7, we can rewrite the condition in equation 9 as:

$$c - \theta > 2(\theta m(s + d))^{\frac{1}{2}} \qquad (10)$$

Firms will choose to deter "cheating" only if the costs of cheating to the firm exceed the benefits of cheating to the worker. That "cheating" entails a net loss is necessary, but not sufficient, for firms to prohibit cheating results from the fact that it is costly to deter cheating. The firm's costs of deterrence include not only monitoring costs, but also the cost of paying wages that are higher than is neces-

sary to attract workers to the firm. In this model it is necessary that workers strictly prefer to work in firms wherein cheating is deterred, otherwise the threat of being fired would impose no costs on a cheating worker, for he could otherwise immediately accept an equally good job in a firm wherein no opportunities for cheating exist.

"Cheating," in quotation marks, can be interpreted as any action a worker can choose to perform which benefits himself but which results in a net loss to worker and firm combined. One example is the consumption of leisure while at work, which is worth less to the worker than it costs the employer, because it is inframarginal leisure or because it is consumed at an inappropriate time and place. Another example may be the worker's consumption or private sale of the firm's products, when these are worth less to the worker than to the firm. Not all actions that result in a net loss are prohibited by the firm. The free consumption of beer by brewery workers obviously results in a net loss, since workers will consume until the marginal utility is zero. The fact that it is sometimes permitted demonstrates that it is costly for the firm to deter "cheating."

The bad jobs in the model are jobs in firms which decide to permit "cheating" as a legitimate emolument which is part of the wage of the job. Such firms' wages (including such benefits) are on their labour supply curves. All bad jobs will yield the same utility to workers. The good jobs in the model are jobs in firms which decide to deter cheating. Such firms' wages will be above their labour supply curves. They could cut wages and still attract workers, but the savings from lower wages would be outweighed by the costs of cheating. Wages are set by firms to maximise profits and not by auctioneers to clear markets. In this model, some firms choose to pay wages above the supply price of labour, and to

ration supply. Not all jobs have equal net advantage. Some good jobs, those with higher gains from cheating, higher monitoring costs, and higher rates of dismissal, will pay higher wages than other good jobs. All good jobs, those wherein cheating is deterred, are better than all bad jobs, where "cheating" is permitted.

One final point may need some clarification. Since all workers are identical, and the firm knows this, why should it fire a worker who has been detected cheating merely to replace him with another worker who is equally likely to cheat in future? The answer is that the firm follows a rule of firing workers whose cheating is detected, in order to maintain workers' expectations that if they are detected cheating then they will be fired. The firm gains no benefit from firing a cheating worker except that it must do so in order to maintain workers' expectations that that same rule will be followed in future, which expectations must be maintained if cheating is to be deterred. The rationality of the actions and expectations that surround deterrence are more fully explored in Rowe (1982).

IV. Unemployment

It is a simple matter to convert the above model of wage differentials into a model of involuntary unemployment. In the previous model we assumed that in equilibrium some firms will choose to permit "cheating." Let us now assume instead that all firms choose to deter cheating. This is not too unreasonable an assumption as most firms proscribe some worker practices which are imperfectly monitored.

All that is really required to convert the model is to reinterpret some of the variables. The wage in bad jobs, v, now becomes the monetary value of leisure plus unemployment compensation, u. The probability per period of switching from a bad job to a good job, s, now becomes the probability of an unemployed worker's being hired, h. The advantages of reinterpreting the model follow not just from the plausibility of the new interpretation, but from the simplicity with which the full market equilibrium can be discussed, now that u can be taken as exogenous.

To best understand the model, it is advantageous to construct the following stability experiment; we consider a hypothetical initial condition, wherein some agents (firms) are not optimizing. We then analyse the process by which equilibrium is attained and unemployment results.

Suppose initially that all firms pay the same level of wages, which is such that the supply of labour exactly equals the demand. In this initial position, all workers will cheat, because even if they are caught and fired, they can immediately get employment in another firm; the penalty is not effective. Suppose one firm realises that it can deter cheating and increase its profits by raising its wages relative to that paid in other firms, since its workers now suffer a fall in income if they are caught cheating and fired. Other firms realize that this stra-

tegy is successful (even if they do not understand why it is successful) and attempt to imitate it by raising their wages relative to that in other firms. Now it is not possible for all firms to raise their wages relative to other firms, but the attempt by each firm to do so means that wages will continue to rise. Wages will not rise indefinitely, because as they rise they will cause excess supply of labour, or unemployment, to rise too. The higher is unemployment, the longer will a fired worker expect to wait before re-employment. Eventually wages and the expected period of unemployment will rise high enough to be sufficient to deter workers from cheating and obviate the need for each firm to attempt to raise its wage above the wage of other firms. Thus equilibrium is reached.[2]

Let us now examine the equilibrium position, wherein wages and unemployment are sufficiently high that no firm need attempt to raise its wages in order to deter cheating. Substituting u and h for v and s in equation 3, we get:

$$\overset{*}{w} = u + \frac{\theta}{p}(h + d) \qquad (11)$$

In full market equilibrium, if we ignore entry to and exit from the labour force, then the number of unemployed workers hired per period must equal the number of employed workers dismissed per period. Letting N be the size of the labour force, and L be the number of workers employed, then:

$$Ld = (N - L)h \tag{12}$$

Rearranging equation 12, we can express the unemployment rate, U, as:

$$U \equiv \frac{(N - L)}{N} = \frac{d}{(h + d)} \tag{13}$$

Letting d be exogenous, and h a market-determined variable, we can solve equations 11 and 13 to express U as a function of w*:

$$U = \frac{d\theta}{p^*(w^* - u)} \tag{14}$$

Equation 14 is not a reduced form solution for the unemployment rate, since *w and p* are endogenous variables. Substituting from equation 7 for p*, the optimum levels of monitoring, we obtain:

$$U = \frac{d\theta m}{(w^* - u)^2} \tag{15}$$

For given levels of d, θ, m and u (assumed exogenous), equation 15 gives us the locus of points in {U,w*} space, such that cheating is just deterred, allowing for an optimal level of monitoring. To complete the model, and obtain a reduced form solution for the unemployment rate, we need two more equations - labour demand and labour supply. We will assume a reverse "L" shaped labour supply curve:[3]

$$N = n \text{ if } w^* > u,\ N = 0 \quad \text{otherwise} \tag{16}$$

Atomistic firms maximize profits taking w^* (the supply price of non-cheating labour) as given. Taking the output good as numeraire, real profits per period, Π, are:

$$\Pi = F(L) - w^*L - mp^*L \tag{17}$$

$$F' > 0 \quad F'' < 0$$

The third term represents the costs per period of monitoring workers.

The profit maximizing condition is:

$$\frac{d\Pi}{dL} = F'(L) - w^* - mp^* = 0 \tag{18}$$

Solving equations 7 and 11 simultaneously, we get the interesting result that optimum monitoring costs per worker per period are equal to the difference between wages and unemployment income:

$$mp^* = w^* - u \tag{19}$$

Substituting 19 into 18 we derive the labour demand curve:

$$F'(L) - 2w^* + u = 0 \qquad (20)$$

Equations 15, 16, and 20, representing the "no-cheating" condition, labour supply, and labour demand respectively, can be solved for full market equilibrium.

$$U^* = \frac{4d\theta m}{(F'(n - nU^*) - \theta)^2} \qquad (21)$$

$$w^* = \frac{u}{2} + \frac{1}{2} F'\left(n - \frac{nd\theta m}{(w^* - u)^2}\right) \qquad (22)$$

Market equilibrium is depicted in Figure One.[4] The quantity of labour employed, and the wage rate, are given by the intersection of the labour demand curve and the "no-cheating" condition. The latter could be thought of as the supply curve of "honest" labour, provided it is remembered that all workers are identical, and are "honest" only if it pays them to be so.

It can be seen that equilibrium will exist with an interior solution if the labour demand curve cuts the "no-cheating" condition in the positive orthant, which will be true provided that:

$$F'(0) > u + (4d\theta m)^{\frac{1}{2}} \qquad (23)$$

We now perform comparative static experiments on equations 21 and 22 to find the effect on the two main endogenous variables, U* and w*, of changes in the six exogenous variables, u, n, d, θ, m, and F(L).

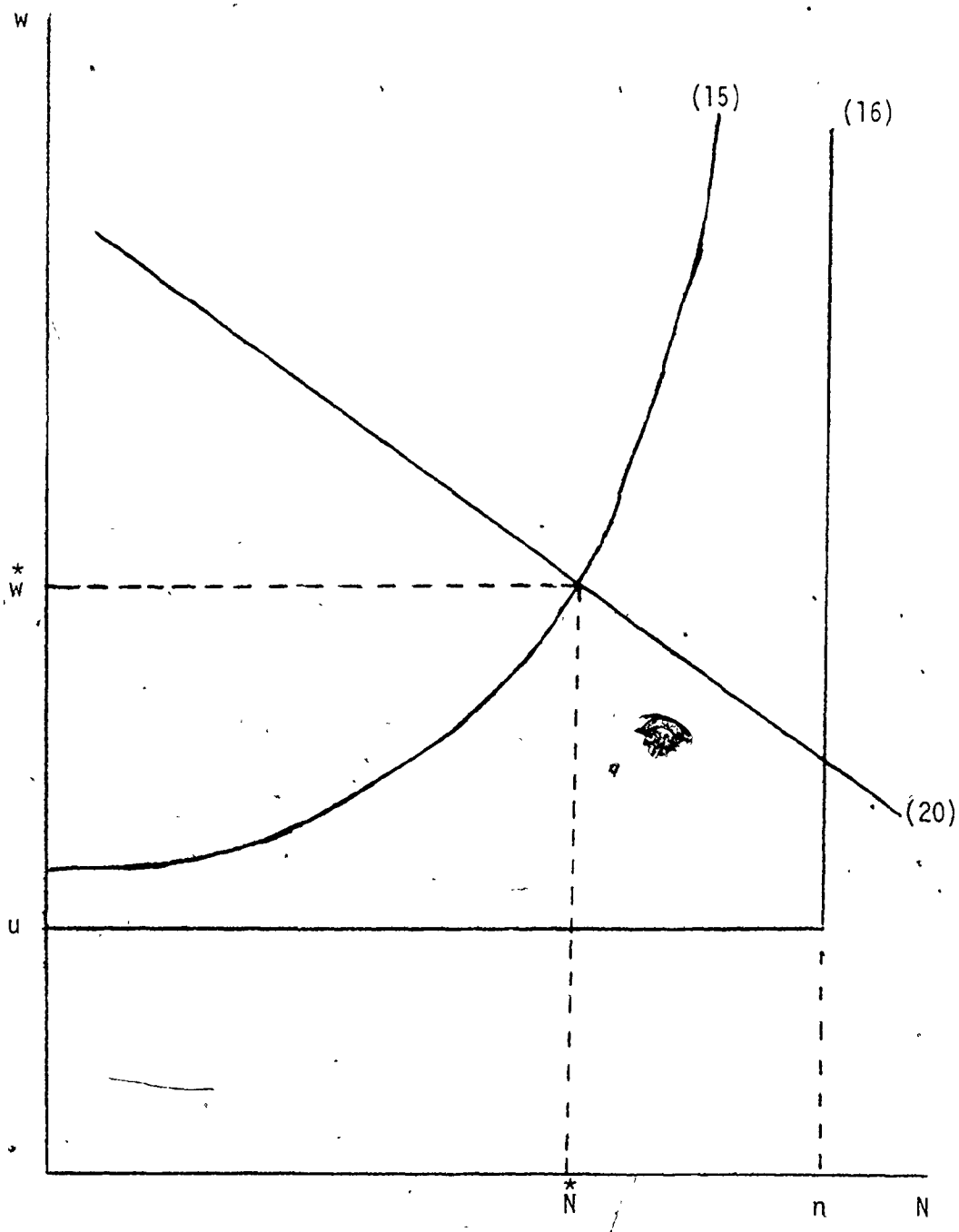


Figure One

$$\frac{d\overset{*}{U}}{du} = \frac{2\overset{*}{U}}{(F' - u) - 2\,nU^{*}F''} > 0 \qquad (24)$$

$$\frac{d\overset{*}{w}}{du} = \frac{(\overset{*}{w} - u) - 2\,LF''}{2(\overset{*}{w} - u) - 2\,LF''} > 0 \qquad (25)$$

An increase in unemployment income will increase wages and unemployment, since it reduces the penalty from being fired, and must be compensated for by a higher wage and longer period of unemployment if cheating is still to be deterred.

$$\frac{d\overset{*}{U}}{dx} = \frac{1/4}{(F' - u)^2 - 2\,nF''(F' - u)} > 0 \qquad (26)$$

$$\frac{d\overset{*}{w}}{dx} = \frac{F''n}{2F''nU^{*}(w^{*} - u)^3 - 2(w^{*} - u)^4} > 0 \qquad (27)$$

where $x \equiv d\theta m$

$$\therefore \quad \frac{d\overset{*}{U}}{dd}, \frac{d\overset{*}{U}}{d\theta}, \frac{d\overset{*}{U}}{dm}, \frac{d\overset{*}{w}}{dd}, \frac{d\overset{*}{w}}{d\theta}, \frac{d\overset{*}{w}}{dm} > 0$$

An increase in the frequency of dismissal, the worker's gains from cheating, or the costs of monitoring will increase the level of wages and unemployment. The first two changes increase the workers' average expected income from cheating relative to not cheating, and must be offset by higher wages and longer unemployment if cheating is still to be deterred. The third change, an increase in monitoring costs, will lead

firms to choose to monitor less frequently and to offset this by increasing the penalty for being caught.

$$\frac{d\overset{*}{U}}{dn} = \frac{-2F''U(1 - U)}{(F' - u) - 2\,nUF''} > 0 \qquad (28)$$

$$\frac{d\overset{*}{w}}{dn} = \frac{(1 - U)\,(w^* - u)F''}{2(w^* - u) - 2n\,(1 - U)F''} < 0 \qquad (29)$$

A rightward shift in the labour supply curve will increase unemployment and lower wages. Wages must fall if more workers are to be employed, and a lower wage reduces the penalty of being fired, which must be compensated for by a higher unemployment rate.

A fall in the marginal product of labour schedule similarly causes a fall in wages and an increase in unemployment. To see this, we introduce a shift parameter 'k,' to multiply the old marginal product schedule $F'(L)$ in equations 21 and 22.

$$\frac{d\overset{*}{U}}{dk} = \frac{2\,UF'}{2UknF'' - (kF' - u)} < 0 \qquad (30)$$

$$\frac{d\overset{*}{w}}{dk} = \frac{(w^* - u)^3F'}{2(w^* - u)^3 - 2\,n\delta\theta mkF''} > 0 \qquad (31)$$

We have seen that this simple model generates twelve comparative static results of unambiguous sign for the effects of changes in the six exogenous variables upon the two main endogenous variables. In principle at least, the model is highly testable.

Unemployment in this model is strictly involuntary. The equilibrium pair of wages and employment is a point off the labour supply curve. Unemployed workers are willing, but unable, to accept jobs immediately at wages less than those paid to identical workers in employment. In contrast to search models of unemployment, for example Mortensen (1970), workers are fully informed about the (real) wages paid by each firm. The reason why the labour market does not clear in this model is that we have placed a restriction on the type of penalty a firm can impose on a worker who is caught cheating. The only penalty a firm can inflict is to discontinue trading with a worker whose cheating is discovered. The cost of such a penalty to the worker depends directly on the wage rate he is paid. If a firm were to reduce its wage towards the market-clearing level, it would have to increase monitoring, or else fail to deter cheating, the costs of either of which would outweigh the gains from a lower wage bill. If we were to relax the assumed constraint on the type of penalty a firm can impose, and allow firms (for instance) to fine or imprison cheating workers, then the link between wages and the size of the penalty is broken, and involuntary unemployment would be eliminated.

In this simple model, unemployed workers are not <u>searching</u>, but <u>waiting</u> for employment. An unemployed worker cannot enter employment until an employed worker leaves his job and becomes unemployed. I have assumed that all unemployed workers face an equal, parametric probability of being hired to fill a new vacancy. In practice, of course, an <u>individual</u> unemployed worker can increase his chances of being hired by actively seeking out those new jobs. However, in this model, in contrast to standard search models, one worker can improve his prospects of getting a job only at the expense of other unemployed workers. If one succeeds in getting a job quicker, this only means that another must wait longer.

Job search is an individually useful but socially useless activity. An increase in aggregate search activity will only reduce aggregate unemployment insofar as search in costly, and increased search will increase the costs per period of being unemployed, which increases the penalty from being fired, and permits lower wages and unemployment. An increase in private search costs per period for unemployed workers is exactly equivalent to a fall in u, the income from unemployment. The same results could be achieved far more efficiently by reducing unemployment insurance. Technological improvements, or subsidies to search activity, if they reduce the private costs per period of job-search, will not reduce but increase the unemployment rate. Anything which improves the welfare of the unemployed can only increase the average duration of unemployment and the unemployment rate, provided, of course, that unemployment results solely from the causes suggested in this essay.

This policy conclusion is extremely pessimistic, moreso since unemployment is strictly involuntary, and individual workers become unemployed through no fault of their own, since all cheating is deterred in equilibrium. Naturally, the policy conclusions of this extremely simple model should be taken with a pinch of salt for practical purposes, but they do warn that attempts to improve the efficacy of search might not have the desired effects.

Eaton and White (1982) assert that, if the need to deter cheating leads to an equilibrium excess supply of labour, then firms can costlessly discriminate between workers along economically irrelevant criteria. If they mean that an individual firm can costlessly discriminate, then their conclusion is true, but is not novel, and has nothing to do with the excess supply of labour. If there is no discrimination in aggregate, then since both groups of workers have the same supply price

an individual firm can hire only (say) white workers with no effect on its profits even in a standard competitive labour market. If, on the contrary interpretation, Eaton and White mean that discrimination is costless even when all firms discriminate, then their assertion is novel, but false. Suppose all firms discriminate against hiring black workers. An unemployed black worker would then face a lower probability per period of being hired. Inspection of equation eleven, however, reveals that a lower value of h for black workers means that the wage needed to prevent black workers from cheating is lower than that for white workers, and therefore discrimination on economically - irrelevant criteria cannot be costless if it is practiced in aggregate. This finding exemplifies the benefits of considering explicitly the flows of workers into and out of employment.

V. Wage Differentials and Unemployment

In the previous section we assumed that condition 10 held for all firms, so that all firms in the economy would choose to adopt the high-wage strategy in order to deter cheating, and used this assumption to construct a model in which involuntary unemployment was present in equilibrium. In this section we will consider how far it is possible to relax this extreme assumption and yet still derive an unemployment equilibrium.

One obvious way to relax this assumption and still get an unemployment equilibrium is to relax the assumption that all members of the labour force are identical. If we divide the labour force into various non-competing sections, perhaps according to skills or location, and if we assume that all employers of a particular section of the labour force will choose to pay high wages to deter cheating, then we can still get involuntary unemployment within that section of the labour force. However, this way of relaxing the extreme assumption is without theoretical interest, for it simply amounts to a redefinition of the labour market, but it may nevertheless be not without practical importance. Firms employing workers with similar skills will be more likely to have similar values for θ, c, d, and m, and so be more likely to make the same decision on whether or not to adopt the high-wage strategy to deter cheating. To segment the labour force is not to relax the extreme assumption that all firms choose to deter cheating, but it does make this assumption more plausible.

We can, however, genuinely relax the assumption that all firms in the labour market choose to deter cheating, and still derive an unemployment equilibrium provided that the demand for labour by firms which per-

mit "cheating" is small enough. More specifically, we need assume only that the quantity of labour demanded by such firms, at a wage equal to unemployment income "u," is less than the number of workers left unemployed by high-wage firms. In equilibrium, an unemployed worker can choose between remaining unemployed and waiting for a "good" job, or accepting a "bad" job immediately, at the same level of income while he waits, and waiting the same expected length of time for a "good" job. If the demand for labour by firms which permit "cheating" is sufficiently small, both wage-differentials and unemployment may coexist in equilibrium.

It is a semantic question whether or not we should call such unemployment "voluntary." A worker's not being employed in a good job is involuntary. His not being employed at all is voluntary. Whether or not we decide to call such unemployment "involuntary" it does bear some resemblance to the results of casual observation; it is rarely the case that an unemployed worker can find no work at all, though the jobs he can get immediately offer inferior wages and/or working conditions than the jobs currently held by his equally qualified peers.

As the model stands, we can generate an unemployment equilibrium even if we relax the assumption that the demand for labour in "bad" jobs is zero, provided that we instead assume that this demand for labour is suitably "small." Some workers will remain unemployed because the utility of being unemployed is equal to the utility of being employed in a "bad" job. If we wish to relax this assumption still further, and derive an unemployment equilibrium with no restrictions on the size of demand for labour in "bad" jobs, we must explain why unemployed workers should not accept an immediate offer of a "bad" job, while waiting for an eventual offer of a "good" job.

An exactly similar problem arises in search theories of unemployment. Why doesn't an unemployed worker accept his first job offer immediately, and work at that job while simultaneously seaching for a better one? The answer given is that job-search is a time-intensive activity, so that an unemployed worker can find a wage offer above his reservation wage more quickly than an employed worker (see Mortensen, (1970)). If this assumption is deemed acceptable, it can be incorporated into the present model. Although search is a socially useless activity in this model, this does not preclude its being useful to an individual, enabling the searcher to wait a shorter period for a "good" job offer. If we assume that an unemployed worker has a higher probability of receiving a "good" job offer than an employed worker, then an unemployed worker may choose to remain unemployed even if his income from a "bad" job is higher than unemployment income, because he expects to wait a shorter time to get a "good" job if he remains unemployed. In other words, firms which permit "cheating" must pay a wage above the value of leisure and unemployment insurance in order to compensate workers for a reduced probability of getting a good job, and attracting them out of unemployment. The necessity of paying higher wages to compensate a worker for having to wait longer reduces the quantity of labour demanded in "bad" jobs, and so increases the equilibrium level of unemployment associated with a given demand curve for labour in such jobs. To restate this conclusion in terms of search-theory, the wage differential between firms that do and do not deter cheating increases the dispersion of wage offers to an unemployed worker, and increases the (individual's) benefits from search.

The assumption used in search theory - that job search is a time intensive activity - is not uncontroversial. It can be argued that unemployed workers spend very little time actively engaged in search, and

that most workers find new jobs before quitting their old one, rather than vice versa. Most "search activity" simply consists of waiting for a suitable job offer to be advertised, which waiting can be performed equally well while employed as well as unemployed.

If the assumption that search is a time-intensive activity is rejected, then we must provide an alternative explanation of why unemployed workers do not accept the first job-offer they receive and work at that job at the same time as waiting for a better offer. One possible explanation is to posit the existence of significant fixed hiring costs - lump sum costs of firm-specific training and relocation, etc., which must be borne by either the worker and/or the firm whenever a new worker is hired. The presence of these fixed hiring costs entails that the joint advantages to firm and worker of initiating an employment relation depend upon the expected duration of that employment relation. It is inefficient for worker and firm to bear these fixed hiring costs if it is expected that the worker will shortly quit.

Suppose that the worker pays some of these fixed hiring costs. The lower is the unemployment rate, the sooner he expects to receive an offer of a "good" job, and the higher the current wage needed to attract the worker into a temporary job. Suppose that the firm pays some of these fixed hiring costs. The lower is the unemployment rate, the sooner the firm expects the worker to receive an offer of a "good" job and quit, and the lower the wage needed to induce the firm to hire the temporary worker. Thus both the supply and demand of labour for "bad" jobs (in firms which do not adopt the high wage strategy to deter cheating) will fall as the expected duration of unemployment falls.

The assumption of fixed hiring costs plays the same role as the assumption that search is a time intensive activity. Both assumptions

make it more costly for the unemployed worker to accept an immediate offer of a bad job while waiting for a better job offer. Both assumptions thus raise the supply price of labour to "bad" jobs above the level of unemployment income, and so reduce the quantity of labour demanded by such firms to a level insufficient to employ all those workers who are not employed in "good" jobs, and thus making it more likely that unemployment will exist in equilibrium.

A summary of the argument of this section is that the assumption that all firms adopt the high-wage strategy to deter cheating is sufficient, but not necessary to generate an unemployment equilibrium. If we relax this extreme assumption we get an equilibrium with wage differentials, which may or may not have unemployment, depending on the size of the demand for labour by firms which permit "cheating." Time intensive job search or fixed hiring costs increase the supply price of labour to firms which "permit" cheating, and thus make unemployment more likely.

VI. The Period of Employment

We have so far assumed that a worker who does not cheat faces a fixed possibility of dismissal equal to "d" per period. However long a worker has remained in his present job, he expects to remain in that job for a further 1/d periods if he does not cheat. The smaller is 'd,' the longer the worker expects to remain in his present job if he does not cheat, and so the lower the wage needed to just deter him from cheating. The role played by this parameter is to discount the future benefits of not being fired for cheating, by the probability of not receiving those benefits because of dismissal for other reasons.[5]

Whereas we have previously taken the period of employment (1/d) to be exogenous, in this section we recognise that it can be endogenous - a choice variable to the firm. The firm may decide to lay-off workers in the event that the marginal value product of labour falls below the wage. The firm realises, however, that the wage needed to deter cheating varies inversely with the expected period of employment as is shown by equation 3. Could the firm increase its long-run profits by guaranteeing greater permanence of employment and thus being able to reduce wages?

It cannot pay to maintain employment in the face of a permanent fall in the marginal value product of labour below the wage rate, but it may pay if the fall is expected to be temporary. It will always pay the firm to guarantee that laid-off workers be given first priority for jobs when rehiring eventually occurs. This guarantee costs the firm nothing, but increases the value to the worker of remaining employed at a given wage. Thus the firm can lower wages and still deter cheating if it makes this guarantee, and so can thereby increase it's profits. Suppose then that a firm lays off workers if their marginal value product falls below the

wage rate, guarantees them first priority in rehiring, and sets wages so that the value of this contract to the worker is just sufficient to deter cheating. There is another strategy which dominates this strategy. While laid-off, the worker will earn the outside wage v (or u if he is unemployed). Thus he would be willing to continue working for any wage at least as great as v. If the marginal value product of labour falls temporarily but remains above v, then there are joint benefits to be gained if employment continues. If the firm maintains employment during such times, these joint benefits allow it to maintain the value of the contract to the worker at a level sufficient to deter cheating, and increase its profits at the same time. If the marginal value product of labour falls below v, then there are no joint benefits from maintaining employment, so the firm will lay-off workers.

Therefore, we would expect to see employment maintained during small and temporary falls in demand, we would expect to see temporary lay-offs during large temporary falls in demand, and permanent redundancies for permanent falls in demand.

There are other devices a firm might use to reduce the wage-premium needed to deter cheating. If all firms that deter cheating could get together and agree never to hire a worker who had been caught cheating by any of their number, then the wage premium needed to deter cheating could be reduced, for a worker fired for cheating would lose not only his present job, but would also destroy his prospects of ever getting another high-wage job. If all firms adopt this strategy of refusing to hire a worker without a letter of reference from his previous employer, then "s" or "h" in equations three or eleven, respectively, are thereby reduced to zero for a worker caught cheating. This reduces the wage needed to deter

cheating, but does not eliminate the wage premium unless "d" is also zero.

The success of this strategy depends on each firm's trusting all other firms to devote resources to scrutinizing each prospective worker's employment record, and never to hire a worker who was once fired for cheating. The individual firm has little incentive to do this, since nearly all the benefits of its doing so will be reaped by other firms. These costs can be reduced if separate firms are set up to specialise in researching and recording a worker's employment record, and to sell this information to firms wishing to hire workers. To ensure that the research firm does not cheat, it agrees partially to insure the employer against losses due to a worker's cheating. In other words, the worker is "bonded," but it is the research firm, not the worker, which loses the bond if the worker is fired for cheating. The worker will still need a wage premium to deter him from cheating, but this premium can be lower since the workers knows that if he is caught cheating he will not be bonded in future. Firms will therefore prefer to hire bonded workers, since their wages need not be as high as unbonded workers.

The specialist firms not only reduce the cost of information, due to economies of scale, but eliminate the need for each firm to trust every other firm to refuse to hire a fired worker, when they have little or no incentive to refuse to do so. The firms need only trust the specialist firm to refuse to bond a cheating worker, and the specialist firms have an incentive to refuse in order to protect their own reputations, and thus protect their bonds and their incomes.

VII. The Employment Relation

In previous sections we have uncritically assumed that workers are employed by wage-setting firms. We have not analysed the concept of the employment relation between workers and firms, nor the reason why labour services should be traded within this particular type of institutional structure.

In order to examine the above questions we will consider the following stability experiment. We will posit a hypothetical initial institutional context for the exchange of labour services, and show how and why a different institutional context - that corresponding to an employment relation between worker and firm - would tend to come into being.

Let us suppose that the workers in a particular occupation are initially self-employed. Each worker sets the price of his labour services at whatever level he chooses, and his customers - the buyers of those labour services - are anonymous and free to switch back and forth between different sellers whenever it pleases them to do so. In short, the labour market we posit corresponds more closely to the casual labour market, where trade between a particular buyer and particular seller is sporadic and infrequent, rather than to the usual labour market where a particular worker gets a "job" with a particular firm, will remain at that job for a considerable length of time, and where the buyer (the firm) sets the wage. Why the former type of labour market is rarely seen will now be explained.

Let us assume that all workers are identical, that there are no barriers to entry to prevent workers from moving freely between this and other occupations, which pay a wage (assumed exogenous) of v.

Suppose that the nature of the labour services traded in this occupation is such as to present workers with opportunities to "cheat" - to take action which is imperfectly monitored by customers, and which costs the customer an amount greater than the benefits to the worker. Let us suppose initially that the wage set by workers in the occupation is such as to ensure equality of net advantage between this and other occupations, so that the labour market clears. At such wages, workers will gain from "cheating," since even if "cheating" is detected and customers were to boycott a "cheating" worker, that worker could exit the occupation and immediately get an equally good job elsewhere. Realizing this, customers will expect all workers to "cheat," so that the latter becomes a legitimate perquisite of the sale of labour services in the occupation. Supposing that the welfare losses from "cheating" are considerable, so that condition 10 is met, would it be possible for self-employed workers in this labour market to adopt the same high-wage strategy as did the firms in previous sections of this paper, thereby creating for themselves a visible disincentive to "cheat"?

An individual worker realises that he can gain if he can convince his customers that he will not "cheat." To do this, he must set his wages at such a level that his income will be high enough so that <u>if</u> detected cheating would forever destroy his reputation, then he would have no incentive to cheat. From equation 3, this requires him to set his wage rate such that:

$$\overbrace{we}^{*} \geq v + \frac{\theta d}{p} \qquad (32)$$

We have introduced a new variable "e," $0 < e < 1$, to represent the proportion of the period that the worker is able to sell his labour services, so that "w^*e" represents his income per period. We have assumed for simplicity that an underemployed worker ($e < 1$) receives zero unemployment income, and that a worker who is once detected cheating and boycotted, will never be able to reestablish his reputation ($s = 0$).

The first worker who sets himself higher wages will have no difficulty in being fully employed ($e = 1$) since, by assumption, customers are better off paying the higher wage if it deters cheating. Other workers, noticing the success of the first, will imitate him by raising their wages and refraining from cheating too. Since the net cost of labour (wages minus cheating costs) to the customer is now lower, the quantity of labour demanded in this occupation will not fall, and all those previously employed will still be fully employed in the new "honest" equilibrium. Let us suppose initially that exactly enough new workers enter to satisfy the increase in the quantity of labour demanded. Thus all workers are fully employed ($e = 1$) and all charge the same wage, w^*, which is just sufficient to deter themselves from cheating.

$$w^* = v + \frac{\theta d}{p} \tag{33}$$

However, a problem now arises since the income of this occupation is higher than in other occupations, so that more workers will seek to enter this occupation. Contrary to what might be expected this entry will <u>not</u> lead to a fall in wages, but will lead instead to a rise. Wages are set by workers, not by a hypothetical auctioneer. No worker will set a wage below w^*, since if he did he would reveal himself to have an in-

centive to cheat, and would get no customers unless his wage was well below v. None will charge more than w*, more than other workers charge, for if they did they would become underemployed ($e < 1$) as customers switched to cheaper workers. If they become underemployed their income (as opposed to their wage) will be insufficient to deter cheating, and they will lose their reputation and be forced to exit. Thus the quantity of labour demanded is fixed at that quantity corresponding to w*. Any successful new entrant must necessarily cause the exit of another worker, by making him underemployed. Customers will always shy away from, and thus reduce the employment, of an underemployed worker, since an underemployed worker has an incentive to cheat.

Equilibrium at w* is only sustainable if new entry is never successful - if customers always prefer (ceteris paribus) established workers to new entrants. If any new entrant is successful, his entry must force the exit of an established worker. If the higher wages cause entry, and so cause exit, they must reduce the expected period of employment (1/d) of a worker in this occupation, which means that wages must rise higher still if cheating is to remain deterred, which encourages a greater flow of would-be entrants, which in turn increases exit, reduces the expected period of employment, and raises wages still further.

Whether or not this process of increasing entry and wages leads to a new equilibrium depends primarily on the rate of entry and the rate at which the rate of entry increases as wages in the occupation increase. If the rate of successful entry is slow, and increases slowly with wages, then a stable equilibrium may eventually be reached with a high, but constant, rate of turnover and wages. If the rate of entry increases quickly as wages rise, then the system will be explosive, with continuously increasing wages and turnover. Eventually, if turnover and wages rise too

much then condition 10 will be violated; customers will prefer to permit "cheating" and pay the competitive level of wages. If the system is explosive, or if it generates a stable equilibrium which would violate condition 10, then trust and reputation can never be established, for if it were established it would be destroyed, and the expectation of its future destruction would prevent its emergence. Even if entry were slow enough to permit the existence of an equilibrium where high wages deter cheating, wages would need to be higher than if the rate of successful entry could be reduced.

The problem, which endangers the existence of an equilibrium in which cheating is deterred, is that a wage high enough to deter cheating is higher than the wage which would clear the labour market and not encourage entry. If the "honest" equilibrium is to be sustained there must be some way to prevent existing workers from being displaced by new entrants. It is not the desire to enter, however, but the effect successful entry may have on turnover which can destroy the "honest" equilibrium. If customers had just the slightest preference for existing workers over new entrants, other things equal, this could prevent existing workers from being displaced by new prospective entrants, and would preserve equilibrium. Alternatively, collusion to restrict entry could benefit both existing workers and customers, since the welfare losses from cheating can be prevented thereby. The problem with this solution is that it may be costly for all existing workers and customers to collude to prevent entry into the occupation as a whole; too many people must agree to take collective action to promote their joint interests. Under certain circumstances, however, it is not necessary that they all collude to set up barriers to entry to the occupation as a whole.

Suppose that a particular customer will need the full-time services of one or more workers for an appreciable period of time. The customer can simply approach the required number of workers, guarantee that he will not hire the services of any other worker unless the original workers are already fully employed, and asks that he be allowed to set the wage rate. Since the workers can always quit if the wage he sets is less than they could get elsewhere, they will not be uninterested in such an offer. A fortiori, since the customer will set a wage above v in order to deter cheating they are positively favourable to such an offer. The customer's guarantee that a worker will get priority of employment is only valid if the worker is not detected cheating. Indeed the customer warns that he will definitely terminate employment in such an event.

What, in effect, the customer who offers such a guarantee has done is to increase the expected period of employment to the one or more workers that he hires. By doing so he has reduced the parameter "d," and thus made it possible to deter those workers from cheating at a finite wage, and at a wage lower than if he had switched his custom from one worker to another at random. In effect, he has demarcated a certain portion of the labour force in that occupation, within which portion he has the power to restrict entry simply by refusing to hire an additional worker if his doing so would reduce the employment (cause the exit) of an existing worker. Provided that an individual customer will require the services of at least one worker for a significant period of time, he does not need to collude with all customers and existing workers in order to restrict entry, and ensure thereby a greater permanence of employment. A further advantage of this institutional arrangement is that the customer need not incur the costs of discovering whether a worker has been

detected cheating by another current customer, for his workers have no other current customer.[6]

Klein and Leffler's (1981) analysis of high prices as a means to deter malfeasance of firms in a product market context gives rise to a similar problem: the level of firms' rents needed to assure a high level of product quality implies that firm's earn economic profits, which provides an incentive for entry of new firms. Since profit-dissipating competition must not diminish the rents needed to assure quality, Klein and Leffler argue that this competition involves firm-specific capital expenditures. These sunk costs serve as hostages to prevent firms from cheating, since a firm detected cheating goes out of business and loses the value of firm-specific assets. Since their analysis lacks anything corresponding to turnover, however, Klein and Leffler are unable to explain what would happen if, for some reason, firms were unable to dissipate profits by firm-specific expenditures. As we have seen, it is the effect of successful entry on turnover which may destroy the "honest" equilibrium. If, as in the labour market, customers give preference to existing suppliers, then the ability of suppliers to yield hostages is not necessary to preserve equilibrium. Positive economic profits are compatible with equilibrium, if those who desire to enter do not actually succeed in taking customers away from existing suppliers. If, for technological reasons, costs cannot be sunk, or if firms fear that yielding hostages makes them vulnerable to blackmail, then positive economic profits will not be dissipated.

An objection might be raised to this theory of the employment relation, that since there is inequality of net advantage, and hence excess supply of labour to this occupation, frustrated workers would have an incentive to set up their own firms, and pay to themselves the high level

wages earned in this occupation. All this objection means, however, is that whenever it is costly for the buyer of a service to prevent the seller from cheating, there is an incentive for buyer and seller to combine into one person. This is perfectly valid. Some people repair their own cars to avoid the high cost of risk of cheating if they pay someone else to do it, <u>despite</u> their having a comparative disadvantage at repairing cars. Some workers purchase or hire land and capital, and sell the finished product, the quality of which may be cheaper to monitor than the quality of their labour services. This theory does not rule out such possibilities, rather it predicts them. The scope for such vertical integration is, however, limited by the standard advantages attributed to the division of labour and to exchange. Conversely, the division of labour is limited, not only by the extent of the market, but also by the extent of trust.

VIII. Relaxing the Key Assumption

The key assumption, on which the novel results of the preceding models depend, is the assumption that the only penalty a firm can inflict on a cheating worker is to dismiss that worker. It is this assumption which forges the link between level of wages and the worker's incentive to cheat. If that link is broken then firms will wish to lower wages, and will thereby eliminate the excess supply of labour, and the novel predictions of equilibrium unemployment or inequality of net advantage will disappear. If, for instance, firms could fine, or imprison, or in any other way impose unlimited costs on a cheating worker which are independent of the level of wages, then firms could, and would, lower wages to eliminate excess supply of labour whilst maintaining a sufficient deterrent against cheating. If the key assumption is dropped, and such penalties alone are used to deter cheating, then the models revert to simple, standard, "text-book" models of the labour market.

The assumption that the only penalty a firm may impose on a cheating worker is to fire him is, perhaps, too extreme and unrealistic. However, the opposite assumption, that firms may freely and credibly threaten other penalties of unlimited size, is equally extreme and unrealistic. What is shown in this section is that the key assumption can be partially relaxed, by allowing the use of other penalties of limited size, without necessarily eliminating the excess supply of labour. The predictions of the models in previous sections are thus weakened, but not necessarily negated.

There are two ways in which we could interpret the limit on the size of penalty that may be imposed. The legal system may prohibit firms from imposing penalties themselves. Any penalties imposed must be imposed by

the legal system, which chooses the size of penalty according to its own standards, independent of what firm and workers may choose to specify in their contract. The limit on the size of the penalty is thus exogenous to firm and worker. A second interpretation is that the penalty takes the form of the forfeiture of a worker's bond. Workers cannot borrow to post such a bond, for if they did, they would have an incentive to cheat and declare bankruptcy, thus the lender, not the worker, would be punished. If workers could be deterred from declaring bankruptcy, those same penalties could be used directly to deter him from cheating. The size of the bond a worker can post, and hence the size of the penalty that can be imposed on him, is thus limited by the size of the worker's non-human wealth. Thus limited liability, or colateral constraints, provides a second interpretation of the limit on the size of penalty which may be imposed.

Let Z designate the value to the worker of the fine (or non-monetary penalty) which is imposed on a worker whose cheating is detected. If the worker follows a strategy of cheating, we must deduct the average expected costs of fines from the expected average income of a worker who follows that strategy. Thus modified, equation 2 becomes:

$$C = \frac{(w + \theta - pZ)s + v(d + p)}{s + d + p} \qquad (2')$$

Setting equal the expected average incomes from the cheating and non- cheating strategies, we solve for the minimum level of wages sufficient to deter cheating.

$$w^* > v + \frac{\theta}{p}(s + d) - Z(s + d) \qquad (3')$$

The larger is Z (the value of the fine) the lower is the minimum level of wages sufficient to deter cheating.

Introducing the fine complicates the firm's decision problem. Whereas before there were only two possible regimes there are now three. In regime A the firm decides to deter cheating and to supplement the fine by setting wages above the supply price of labour. In regime B the firm decides to deter cheating, set wages equal to v, and rely on the fine alone. In regime C the firm decides to permit "cheating" and sets wages equal to $(v - \theta)$.

Consider regime A. The firm seeks to minimize the costs of wages plus monitoring costs per worker.

$$M + w^* = mp + v + \frac{\theta}{p}(s + d) - Z(s + d) \qquad (5')$$

Provided it remains in regime A, the optimal level of monitoring, p_A, is independent of the size of the fine.

$$p_A = \left(\theta(s + d)/m\right)^{\frac{1}{2}} \qquad (7')$$

Substituting 7' into 3', the optimum level of wages in regime A is:

$$w_A = v + \left(\theta(s + d)m\right)^{\frac{1}{2}} - Z(s + d) \qquad (33)$$

Since the firm cannot set wages below the supply price of labour, if the solution to 33 implies that $w_A < v$, then the firm will not be in regime A.

Consider regime B. The firm sets wages equal to the supply price of labour and chooses a level of monitoring such that the fine will just deter cheating.

$$w_B = v \ , \ p_B = \frac{\theta}{Z} \tag{34}$$

In regime C the firm sets wages equal to the supply price of labour and does not monitor "cheating."

$$w_C = v - \theta \ , \ p_C = 0 \tag{35}$$

Now consider the firm's choice between regimes B and C. The firm will prefer regime B to C if and only if the costs of wages plus monitoring in B are less than the costs of wages plus cheating in C:

$$B > C \quad \text{iff} \quad w_B + mp_B < w_C + c \tag{36}$$

which implies, given 34 and 35:

$$B > C \quad \text{iff} \quad Z > \theta m/(c - \theta) \tag{37}$$

The firm will choose A rather than C if and only if the wage premium plus monitoring costs is less than cheating costs minus the worker's gains from cheating.

$$A > C \ \text{iff} \ mp_A + w_A < v - \theta + c \tag{38}$$

which implies, given 7' and 33:

$$A \succ C \text{ iff } Z > 2(\theta m/(s + d))^{\frac{1}{2}} - (c - \theta)/(s + d) \qquad (39)$$

To complete the choice of regime, we note that A is not feasible if the wage in A would be below the supply price of labour, and that A dominates B otherwise:

$$A \succ B \text{ iff } w_A > v \qquad (40)$$

which implies, given 33:

$$A \succ B \text{ iff } Z < (\theta m/(s + d))^{\frac{1}{2}} \qquad (41)$$

The most important result of this section is that introducing the possibility of limited fines does not necessarily eliminate the use of excess labour supply as an <u>additional</u> deterrent against cheating. The firm's costs of thus increasing the penalty may be offset by the reduction in monitoring costs it permits. However, if the firm remains in regime A, the fine does reduce the wage rate and hence reduces the excess labour supply. Introducing the fine reduces the firm's costs of deterring cheating, and thus makes it more likely that the firm will move from regime C into either regime A or B. It is thus possible, therefore, that the introduction of a limited fine could create an excess supply of labour where there was none before. A limited fine might increase wages above the supply price of labour. For a firm already in regime A, however, increasing the level of the fine reduces the wage and excess labour

supply until eventually the latter is eliminated and the firm adopts regime B.

IX. Conclusion

The purpose of this paper has been to analyse the implications of the assumption that the only incentive an agent has to deliver on an exchange is to maintain his reputation for doing so. Since the value of his reputation depends on the price and quantity of current and future expected exchanges, the attempt to modify incentives to deliver can have an effect on prices and quantities traded. In the setting of an atomistic labour market we show that this assumption may lead to an equilibrium excess supply of labour to a particular firm. This excess supply of labour may be revealed as wage-differentials, unemployment or both. We further analyse the implications of this assumption for the institutional setting within which the exchange of labour services is conducted. Relaxing the key assumption by introducing other penalties of limited size weakens, but need not nullify, the predictions of the models.

Although we have chosen a labour market as the setting for the models, no restriction on the scope of their applicability was thus intended. These models (or some variant thereon) are applicable in any market where the incentive for honouring an exchange contract (explicit or implicit) depends in whole or in part on maintaining one's reputation for doing so. We need not even assume that the quality of the good traded is a choice variable for one of the parties to the exchange. "Cheating" may equally well be interpreted as lying about the quality of the good one is either buying or selling, which quality may be exogenous for both parties. What is important is that whatever action is interpreted as "cheating" should impose welfare losses. Lying about the quality of a good traded will result in net welfare losses if it leads to exchanges taking place which are not mutually beneficial at some price.

This may be true if, inter alia, buyer and seller evaluate quality differently. For example, we could apply this model to the "Market for Lemons." Unlike Akerlof (1970), however, we would not impose the assumption that the price must clear the market for used cars. In that respect our model resembles that of Weiss (1980) who also shows that rational price-setting behaviour may conflict with price-setting to clear markets. Unlike Weiss, however, the link between price and quality operates through the reputation for honestly revealing quality, and not via the effect of quality on the value of refraining from selling and hence on the supply price. The general conclusion of this paper is that for any problem, in particular for any principal-agent problem, where it is costly or impossible to enforce an optimal contract, it is not legitimate simply to impose the assumption that prices must clear markets. If prices do clear markets, this must be shown to be the result of rational price setting behaviour by profit maximising firms or utility maximising agents. One example will suffice to illustrate this point:

Consider a risk neutral farmer and a risk neutral landowner. Output depends on the farmer's effort and on a stochastic variable, neither of which is observable to the landowner. The optimal contract would require the farmer to pay a fixed rent for the land, for otherwise he will not receive the full marginal expected return to his effort and will undersupply effort. This optimal contract may, however, be infeasible if the farmer has no wealth and the output under the worst outcome falls short of the market clearing rent. Rather than adopting an inefficient contract with an output-contingent rent (e.g., sharecropping) it <u>may</u> pay the landowner to retain a fixed rent but to reduce the rent below the market-clearing level. In other words, since the efficiency of the optimal feasible contract depends on the farmer's wealth, it may pay the landlord

to give the farmer more wealth. Even the assumption of limited liability is not strictly necessary. A risk neutral principal might benefit by giving wealth to his risk averse agent in order thereby to promote a greater similarity in their evaluation of the probability distribution of returns from any action by the agent. In short, all models of the principal agent problem wherein market clearing is simply imposed should be considered guilty of ad hocery until proven innocent. Some will not be proven innocent. The supply price of agents imposes only a lower bound on the value of their remuneration, not an equality.

The Coase Theorem asserts that if rational agents perceive that an exchange would be mutually advantageous, then that exchange will take place, and the exhaustion of all mutualy advantageous exchanges implies that a Pareto Optimal allocation must be attained. This is palpably false, as any reader of crime and spy novels will know, if agents cannot trust that the terms of an exchange contract will not be violated. It is, of course, possible to save the Coase Theorem by invoking "transactions costs," but to do so is merely to put a name on one's ignorance. One application of the Coase Theorem concerns not exchanges within the law, but the exchange of laws. Rational agents will not permit an inefficient legal system to persist.

If firms and workers can costlessly contract for unlimited penalties to be imposed by the legal system on a cheating worker, then (in this model at least) they will choose to do so, and reputation will be a redundant and expensive way to enforce contracts. We have already considered several reasons why firms and workers might not be able to contract costlessly with the legal system for unlimited penalties: it may be costly to make explicit an implicit contract. It may be costly to prove that cheating occurred. Proof and explicit contracts will be

required if workers do not trust firms to resist their temptation to blackmail workers. Even if workers do trust firms, the legal system may fail to support, or may even prohibit, the imposition of penalties by firms on workers because the legal authorities may not trust firms to refrain from abusing such powers. Yet even if we ignore these plausible reasons why firms and workers might not be able costlessly to contract with the legal system for unlimited penalties, there remains a further reason why it should not be so, which reason is internal to the model at hand. The model at hand shows that there is excess supply - that some mutually advantageous exchanges will not be made - in the absence of a third party to enforce the terms of an exchange contract. But then when we consider the mutually advantageous exchange of one set of laws for another - when we consider those exchanges that define or constitute the legal system - we recognize that no such third party exists, for he is one of the trading partners. The police cannot post bonds, as in Becker and Stigler, for who is to decide whether those bonds are to be forfeited? In a general equilibrium view of the enforcement of exchange contracts, there is no third party, the sufficiency of whose incentives honestly and diligently to enforce the terms of exchange can be assumed. All exchanges are ultimately enforced by reputation, and by reputation alone. The reputation that is relied on can be shifted from one agent to another, but it is only the reputation of someone, somewhere, which enforces exchange and makes exchange possible.

Footnotes

* I would like to thank David Laidler, Chris Robinson and Peter Howitt for commenting on earlier versions of this paper. I am grateful to Tim Lane for long discussions which helped to clarify and develop the ideas underlying this paper. Any shortcomings are my own responsibility. The first typed draft of this essay is dated August 1980. This essay was written independently of Shapiro and Stiglitz (1984), which covers similar topics though in less detail. Another recent paper, Yellen (1984), also surveys models of unemployment, including the "shirking model."

1. We assume that the period is sufficiently short that the product dp, the probability of being simultaneously dismissed and fired in the same period, is zero.

2. It is important to realize that we are not proposing a conspiracy theory whereby firms deliberately create unemployment in order to discipline workers. An individual firm cannot affect the rate of unemployment, and so cannot intend to do so. Instead, we are proposing the theory that unemployment is the <u>unintended consequence</u> of the attempt by each firm to raise its wage above other firms in order to deter cheating. Unemployment results from the attempts by firms to do something which, in aggregate, they cannot do.

3. An upward sloping labour supply curve would introduce complexity, since u would vary across workers, and the average level of u for those in the labour force would become endogenous.

4. The positions of the three curves are not independent. The height of the demand curve depends on u, and any move in the supply curve will shift the "no-cheating" condition in a similar direction.

5. There is thus no need to interpret this parameter solely as the probability of dismissal; it could also be interpreted as an interest rate or as the probability that a worker will choose to quit a particular job for extraneous reasons.

6. I do not claim this theory to be a complete account of the nature and existence of firms. It does, however, provide an explanation of the nature and existence of the employment relation between firms and workers, whereby existing workers take priority over jobs, and each worker has only one employer at a point in time. The employment relation between workers and firm is part of that set of implicit and explicit contractual relations which together constitute, or which we designate as, "the firm."

Appendix

I. Adam Smith on Trust and Wage Differentials

The original statement of the argument that wages depend on trust is found, not in Adam Smith, but in Richard Cantillon, from whom Smith is known to have borrowed.[1] Cantillon's Essai, however, contains just the single sentence: "When capacity and trustworthiness are needed the labour is paid still more highly, as in the case of jewellers, book keepers, cashiers and others" (p. 21). No explanation is given as to why the need for trust might be associated with higher wages. Smith, unlike Cantillon, does give some words of explanation, but it is not easy to give a consistent interpretation of Smith's explanation.

"We trust our health to the physician, our fortune, and sometimes our life and reputation, to the lawyer and attorney. Such confidence could not safely be reposed in people of a very mean or low condition. Their reward must be such, therefore, as may give them that rank in the society which so important a trust requires" (Wealth of Nations, p. 105).

If the above quotation had stood alone, it might have been possible to have given a consistent interpretation of Smith's explanation along the lines of the explanation presented in this paper. High wages, or the rank in society which they can buy, gives the physician or lawyer something to lose if he is revealed to be in breach of trust. They thus provide the appropriate incentive. However, such an interpetation would be inconsistent with Smith's statement at the beginning of the same chapter.

"The five following are the principle circumstances which, so far as I have been able to observe, make up for a small pecuniary gain in some employments, and counter-balance a great one in others: ...fourthly, the small or great trust which must be reposed in those who exercise them... (p. 100).

Smith here clearly states that wage-differentials due to trust do not violate the principle of equal net advantage between employments. What then, we might ask, is the disadvantage which counterbalances the high wages in employments which require trust?

Perhaps we might interpret Smith along the lines of Stigler and Becker (1974). High pay in later years deters malfeasance, but is offset by low or even negative pay in earlier years to ensure the equality of net advantage and eliminate excess supply. Smith, however, does not state that workers in employments requiring trust get low wages in earlier years. Moreover, if the lifetime income of a trusted worker were no higher than that of other workers, it would not suffice to "give them that rank in society which so important a trust requires," nor to raise them from a "very mean or low condition."

If we wish to retain both high lifetime incomes and equality of net advantage, we must assume that undertaking responsibility, or maintaining a high rank in society, is unpleasant. Smith does not state this to be the case, and if he had thought it so he would simply have included trust as a particular example of the first cause of wage differntials - the agreeableness or disagreeableness of the employments themselves. Moreover, if responsibility were unpleasant this would exactly offset the incentive effect of high wages and provide no reason to trust those of a high rank.

We might reject the interpretations that explain high wages according to their incentive effect, and interpret Smith as meaning that high wages are the rents paid to a scarce, natural attribute, trustworthiness. The most trustworthy will enter occupations where trust is most required, and so wages in those occupations will be higher than average. Again, Smith does not state this to be the case, and if it were the case then the principle of equal net advantage between employments would be violated. Furthermore, if Smith had recognized that the wages of employments vary according to this attribute of the average individual in those employments he would surely have recognized, and stated, that differences in other natural attributes can also cause differences in wages. He does not.

Curiously, Smith does recognise that the profits of individuals might vary according to the attribute of trustworthiness of those individuals, but he does not infer from this that the wages of employments might vary according to the attribute of trustworthiness of the average individual in those employments.

"When a person employs only his own stock in trade, there is no trust; and the credit which he may get from other people, depends not on the nature of his trade, but upon their opinion of his fortune, probity, and prudence. The different rates of profit, therefore, in the different branches of trade cannot arise from the different degrees of trust reposed in the traders" (p. 105).

This passage clearly shows that Smith failed to see the possibility that more trustworthy individuals might congregate in trades and employments where more trust was needed.

In the absence of any other proposed interpretation, I conclude that there is no consistent interpretation of Smith's explanation of the relation between trust and wages.

References

Akerlof, G.A. (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." Quarterly Journal of Economics, 84, August, 488-500.

Calvo, G. (1979), "Quasi-Walrasian Theories of Unemployment," American Economic Review, 69, No. 2, 102-107.

Calvo, G. and S. Wellisz (1978), "Supervision, Loss of Control, and the Optimum Size of the Firm." Journal of Political Economy, 86, October, 943-952.

Calvo, G. and S. Wellisz (1979), "Hierarchy, Ability and Income Distribution," Journal of Politicl Economy.

Cantillon, R. (1964), Essai sur la nature de commerce en général. Ed. H. Higgs (1931), reproduced by A.M. Kelly, New York.

Eaton, B.C. and W.D. White, (1982), "Agent Compensation and the Limits of Bonding," Economic Inquiry, 20, No. 3 (July).

Gintis, H. and Bowles, S. (1981), "Structure and Practice in the Labor Theory of Value," The Review of Radical Political Economics 12, No. 4.

Klein, B. (1980), "Borderlines of Law and Economic Theory: Transaction Cost Determinants of 'Unfair' Contractual Arrangements," American Economic Review and Proceedings, 70, May, 356-362.

Klein, B. and K. Leffler (1981), "The Role of Market Forces in Assuring Contractual Performance," Journal of Political Economy, Vol. 89, No. 4, 615-641.

Lazear, Edward P. (1979), "Why is There Mandatory Retirement?" Journal of Political Economy, 87, No. 6, December, 1261-1284.

Mortensen, D.T. (1970), "A Theory of Wage and Employment Dynamics." In Phelps, ed., Microeconomic Foundations of Employment and Inflation Theory. New York: Norton.

Rowe, N. (1982), "The Rationality of Promises," Carleton University, mimeo.

Shapiro, C. and J.E. Stiglitz (1984), "Equilibrium Unemployment as a Worker Discipline Device," American Economic Review, Vol. 74, No. 3, June.

Smith, Adam (1776), The Wealth of Nations. New York: Modern Library, 1937.

Stigler, G. and G. Becker (1974), "Law Enforcement, Malfeasance and Compensation of Enforcers." Journal of Legal Studies 3, No. 1, 1-18.

Weiss, A. (1980) "Job Queues and Layoffs in Labor Markets with Flexible Wages," Journal of Political Economy, 88, No. 3, June, 526-538.

West, E.G. (1980), "Richard Cantillon and Adam Smith: A Reappraisal," Carleton University.

Yellen, J.L. (1984), "Efficiency Wage Models of Unemployment," American Economic Review, Vol. 74, No. 2, May.

Chapter Five

An Equilibrium Theory of Strikes

I. Introduction

A strike is a conflict over the distribution of gains from trade. The two sides withdraw from trade, each demanding a larger share of the joint rents to their relationship than the other is prepared to offer. To the economist, strikes are puzzling simply because they are costly - a portion of the rents is destroyed by the conflict over their distribution. There must exist some distribution such that both sides could gain by accepting that distribution rather than resorting to a strike. Why then do presumably rational agents fail to make just such a mutually beneficial exchange?

This paper puts forward the theory that strikes are a consequence of efficient rules of action rationally adopted by firm and worker to enforce a long-term implicit contract which sets the wage (and possibly other conditions of employment) contingent on a set of "conditioning variables," when the firm and worker have imperfect and partially independent information on the values of those conditioning variables. Suppose that it is efficient for a long-term employment relations to be governed by an implicit contract which indexes the wage to (say) the price level. Since it may be costly to monitor the price level continuously, it will be monitored at discrete intervals (say) one year, with the wage being set fixed during the year. These yearly wage-fixings correspond to the short-term explicit contracts usually observed in the labour market. These short-term explicit contracts are not negotiated in a vacuum, however, but should be seen as interpretations of a long-term implicit contract. In order to enforce the implicit contract, each side must deter violations by following a rule of imposing costs at least

equal to what the other side can expect to gain by violating the contract. The obvious way to do this is to withdraw from trade - returning only when the wage and strike length are such that the rents accruing to the other side are no greater than it could have gained by accepting the contractual wage with no strike. Faced with such a rule of action, neither side will wish to violate the implicit contract. If both sides agree on the contractual wage, no strike will occur. If the price-level is imperfectly observed, however, and the two sides have partially independent information on the price-level, it may sometimes be the case that the worker observes a "high" price-level, requiring a "high" contratual wage, while the firm observes a "low" price-level, requiring a "low" contractual wage. Ex ante, neither side can tell whether the other's announced observation of the price level is honest or whether it is dishonest and merely an attempt to capture a larger share of the rents than is justified by the implicit contract. In order to preserve the incentive for honest announcements, and hence to preserve the implicit contract, each side must act as if the other's announcement were dishonest, and the disputed rents are destroyed by a strike.

If this theory of strikes is valid, then the closest analogy to strikes is perhaps the punishment of criminals. In order to deter violations of the law, the courts must impose costs on convicted criminals sufficient to ensure that the net expected benefits of committing a crime are non-positive. Why then does crime and punishment exist in equilibrium? Why not threaten punishment so draconian that no crimes are ever committed, thus eliminating the social costs of crime and of punishment? One answer might be that monitoring is imperfect, so that innocent people are sometimes convicted.[1] The draconian penalties sufficient to elimin-

ate convictions altogether would impose major costs as people attempted every means to avoid the slightest risk of suspicion.[2]

Strikes, and the punishment of convicted criminals, are activities which are socially costly ex post. Rational agents engage in these activities in order to maintain their reputations for following rules which enforce a contract, the value of maintaining which exceeds the costs of enforcement. Some strikes occur in equilibrium because strike lengths sufficient to deter all disagreements would mean that (say) the firm would never announce a fall in the price level for fear of the small probability that it might thereby provoke a strike, and the wage would no longer vary with the price level.[3]

Unlike existing theories of strikes, the theory presented here is compatible with rationality of action and expectation, and with efficient contracting. In Ashenfelter and Johnson (1969), the workers could gain by instructing their union to accept the firm's initial offer. In Cross (1965), both workers and firms have non-rational expectations about the outcome of a strike. The theories of Crawford (1982) and Hayes (1984) are an advance in that both impose rationality on the players within the context of the game, but in neither case is the mechanism which creates stikes ex ante efficient. In Crawford's theory both sides could gain by agreeing to renounce the use of strategies involving precommitment. In Hayes' theory both sides could gain if the workers were paid a fixed wage before the state of the world is revealed. It is not explained why this does not happen. Since the strike-mechanism in this paper is efficient ex ante, such problematic questions do not arise.

In the next section of this paper we present a simple model to provide an illustration and example of the theory of strikes. Risk aversion on the part of both agents motivates a contract which sets a wage contin-

gent on the firm's product price. The firm observes the true product price, while the worker observes a "price" variable which may be more or less perfectly correlated with the firm's price. At one extreme, with zero correlation, the worker gets no information on the firm's price, and the model is formally analogous to the theories of ex post inefficient layoffs under asymmetric information advance by e.g., Grossman and Hart (1981). At the other extreme, with perfect correlation, the product price becomes public information, and strikes are threatened, but never occur in equilibrium. We examine the properties of the model, and hence of strikes, in the limit as information approaches full publicness.

It should be understood that the model of strikes is presented merely to provide a specific example of how this theory of strikes might be formalized. The theory itself is far more general, and should remain applicable under diverse conditions. The key ingredients are only that there be some motive for a contract setting certian contractual variables contingent on some conditioning variables, where the firm's and worker's estimates of those conditioning variables are correlated but not perfectly correlated.

II. The Model

Assume the labour market to consist of an even number of identical risk-averse agents. These agents form into pairs, one being designated "the firm" and the other "the worker." Each pair carries out production and consumption on one of a number of identical isolated islands.[4] They arrive on the island at the beginning of the period and cannot leave until the end of the period.

Output is produced under constant returns[5] to labour input, N, subject to the resource constraint that $N \leq 1$. N thus represents output, employment, or the fraction of the period during which production takes place. The worker has immediate control over the level of N (he cannot be forced to work).

On arriving at an island, the firm privately observes P_F, the price at which he can sell output for consumption goods. The worker privately observes P_W, a variable which has no significance in itself, but is relevant insofar as it gives the worker more or less accurate information on P_F. The joint probability distribution of P_F and P_W is common knowledge to both agents. The firm sells the output, N, at the price P_F, for an amount of consumption goods NP_F.[6] The consumption good cannot be stored nor transported to or from the island; the sum of firm's and worker's consumption must therefore equal NP_F. The firm pays the worker NW, where W is the (real) wage rate, and consumes the remaining $N(P_F - W)$. The firm has immediate control over the payment of wages (he cannot be forced to pay wages). There is no disutility from labour; each agent's utility thus depends only on his own consumption, with everywhere positive and diminishing marginal utility.

An efficient, and hence equilibrium, contract will specify wages and employment, contingent on firm's and worker's announcements, to maximize the sum of their expected utilities. Notice that there is no need explicitly to impose the market clearing condition of equality between firm's and worker's expected utilities; this can always be achieved by the contract requiring the pair to toss a coin to determine who plays the role of the firm and who the role of worker.[7]

IIa. Public Information

If P_W were identically equal to P_F, then P_F would be public information, and the solution of the equilibrium contract is straightforward. It will set $N = 1$ (full employment) in all states of the world. Total consumption is thus P_F, which must be divided equally between the two identical risk-averse agents to maximize the sum of their expected utilities. This gives a contractual wage of $W = P_F/2$.

Even with public information, however, the contract must still be enforced. Publicness ensures only that both sides agree on what would constitute a violation of the contract - it does not per se prevent violations from occurring. Since the level of employment and output is the worker's immediate decision, how does the firm induce the worker to supply the contractual amount of labour? Similarly, since the payment of wages is the firm's immediate decision, how does the worker induce the firm to pay the contractual level of wages? To deter violations, each side must follow a rule which ensures that the other's consumption cannot exceed the contractual level, $P_F/2$. Setting the worker's consumption, NW, equal to the contractual level, we find the firm's "wage-offer curve":

$$W^o = P_F/2\ N \qquad (1)$$

After a strike of length $S \equiv (1 - N)$, the firm can offer no more than W^o if it wishes to ensure that the worker does not gain by demanding a higher than contractual wage.

Similarly, setting the firm's consumption, $N(P_F - W)$ equal to the contractual level, we find the worker's "wage-demand curve":

$$W^d = P_F - P_F/2N \qquad (2)$$

After a strike of length $S \equiv (1 - N)$ the worker can accept a wage no less than W^d if he wishes to ensure that the firm does not gain by offering a lower than contractual wage.

These two "concession curves" are depicted, as rectangular hyperbolae, in Figure One. These two concession curves enforce the contract, since neither side can do better than accept the other's initial offer of the contractual wage and no strike. Solving the two concession curves simultaneously yields the strike length and eventual wage settlement. In this case, since both sides have the same beliefs about the contractual wage, the solution is a strike of zero length and a settlement wage equal to the contractual wage. This is the result of public information. Just suppose however, that some "accident" resulted in (say) the worker believing P_F to be above its true value, and hence the contractual wage to be above its true value. The intercept of the worker's wage demand curve (which equals his perceived contractual wage) now lies above the intercept of the firm's wage offer curve. The two concession curves now intersect at a positive strike length. What is happening is that

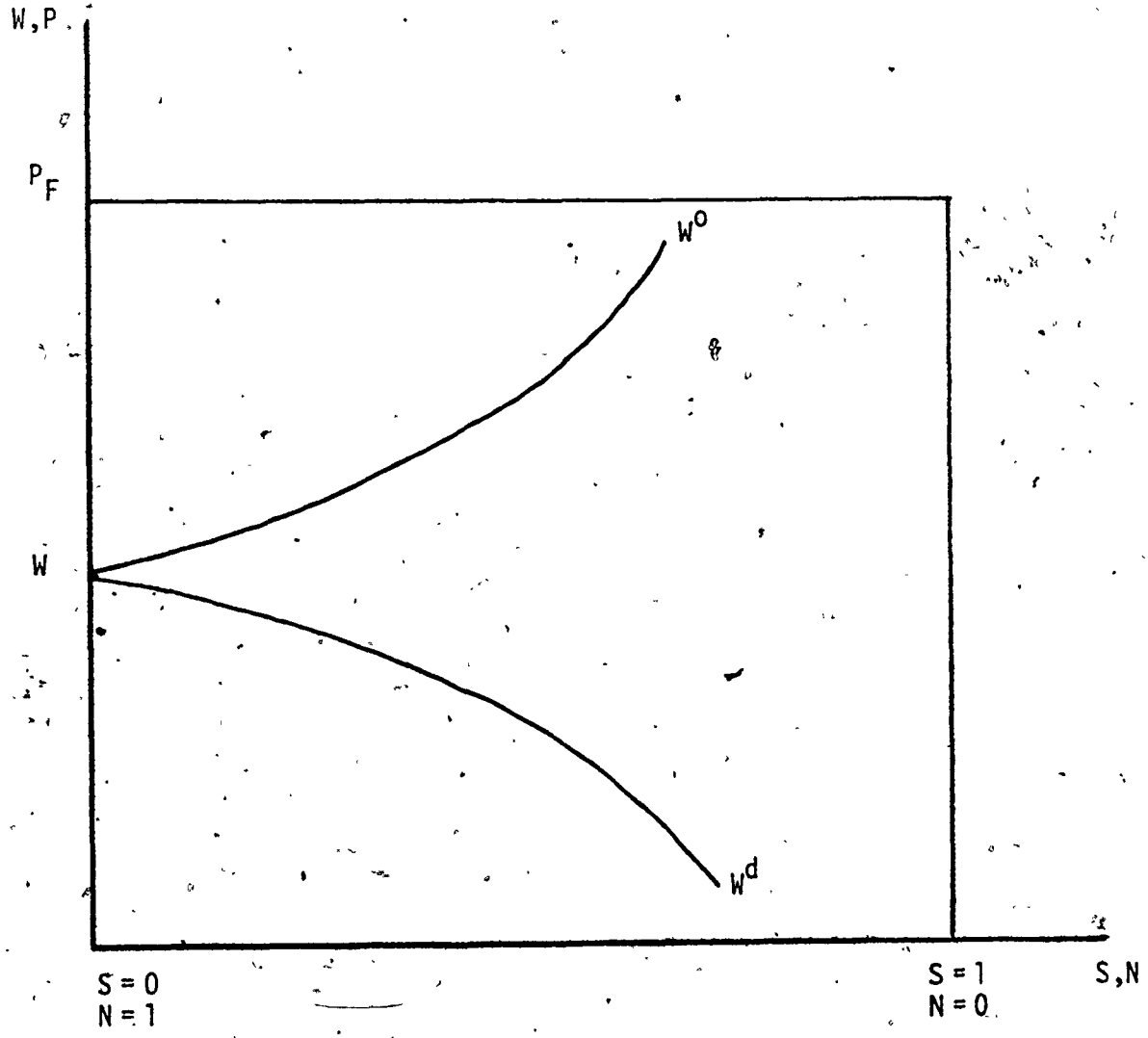


Figure One

each side interprets the other to be attempting to violate the contract by proposing a wage more favourable to itself than the contractual wage. Each punishes the other's perceived violation by refusing to trade at the other's proposed wage, but as the strike proceeds each can concede slowly in proportion to the costs already imposed on the other. The outcome is that each side gets exactly what the other side believes it deserves; the strike destroys rents exactly equal in value to the rents in dispute.

This above example, though useful for illustration, is not strictly legitimate. We have permitted an "accidental" difference in beliefs in a world where agents believe with certainty that such "accidents" cannot happen, thus violating rationality of expectations. To correct this problem, we must formally introduce a framework in which information can be private, thus permitting agents to recognize the possiblity of differences in beliefs.

IIb. Private Information

To proceed further, we must specify the joint probability distribution of P_F and P_W. Assume that each can take one of two values, H (high) and L (low). We thus have four actual states of the world, which we designate by the lower case letters, a, b, m, and ♠.

(a) (H_F, H_W) if $P_F = P_W = H$

(b) (L_F, L_W) if $P_F = P_W = L$

(m) (H_F, L_W) if $P_F = H$, $P_W = L$

(s) (L_F, H_W) if $P_F = L$, $P_W = H$

Letting $\Pi(\cdot)$ designate the prior probability of a state, we will assume high and low prices are equally probable, so that:

$$\Pi(a) = \Pi(b) = \Pi/2 \tag{3}$$

$$\Pi(m) = \Pi(s) = (1 - \Pi)/2 \tag{4}$$

From (3) and (4) we find the conditional probabilities:

$$\Pi(H_F/H_W) = \Pi(H_W/H_F) = \Pi(L_F/L_W) = \Pi(L_W/L_F) = \Pi \tag{5}$$

$$\Pi(H_F/L_W) = \Pi(L_W/H_F) = \Pi(L_F/H_W) = \Pi(H_W/L_F) = (1 - \Pi) \tag{6}$$

The parameter Π thus indicates the probability that firm's and worker's observations will "agree" and $(1 - \Pi)$ the probability they will "disagree." The parameter Π can therefore be understood as an index of publicness of information. At one extreme, if $\Pi = 1$, the worker observing P_W knows P_F with certainty. At the other extreme, with $\Pi = 1/2$, observing P_W gives the worker no information on P_F.[8]

Matching the four actual states are four "announced states," which we designate by the corresponding upper case letters: A, B, M, and S. Thus A is the announced state if both firm and worker announce high observed prices, etc.

A contract specifies wages and employment contingent on the announced state, so that W_A and N_A are the contractual levels of wages and employment if A is the announced state. Since the worker's consump-

tion is equal to the wage times employment, his utility, which we designate by V, depends only on the announced state. The contract therefore specifies four contractual levels of utility for the worker: V_A, V_B, V_M, and V_S. Since the firm's consumption is equal to $(P_F - W)N$, his utility, which we designate by U, depends on both the announced state and on the actual value of P_F. There are thus eight contractual levels of utility for the firm, U_A, U_B, U_M, and U_S if the firm's announcement is true, and $\tilde{U}_A$, $\tilde{U}_B$, $\tilde{U}_M$ and $\tilde{U}_S$ if his announcement is false.[9]

We now consider the constraints imposed on the contract by the requirement that it be truth-revealing, that neither agent have an incentive to make a false announcement. We will assume that announcements must be made simultaneously, so that each makes his own announcement in ignorance of the other's announcement. (It can be verified that sequential announcement cannot dominate simultaneous announcement, since the constraints imposed by the former constitute a subset of the constraints imposed by the latter.)[10]

Consider first the firm's truth-telling constraints. If the firm observes H_F, it knows that the worker will observe H_W with probability Π, and will observe L_W with probability $(1 - \Pi)$. Assuming the worker tells the truth, if the firm tells the truth it will receive U_A with probability Π, and U_M with probability $(1 - \Pi)$; if the firm lies it will receive $\tilde{U}_S$ with probability Π, and $\tilde{U}_B$ with probability $(1 - \Pi)$. For the firm to have no incentive to lie, its expected utility from announcing H_F must not be less than its expected utility from announcing L_F:

$$\Pi U_A + (1 - \Pi) U_M \geq \Pi \tilde{U}_S + (1 - \Pi) \tilde{U}_B \qquad (7)$$

Similarly, for the firm to tell the truth on observing L_F:

$$\Pi U_B + (1 - \Pi) U_S \geq \Pi \hat{U}_M + (1 - \Pi) \hat{U}_A \qquad (8)$$

Similarly, assuming that the firm tells the truth, the worker will tell the truth on observing L_W if:

$$\Pi V_B + (1 - \Pi) V_M \geq \Pi V_S + (1 - \Pi) V_A \qquad (9)$$

and for the worker to tell the truth on observing H_W:

$$\Pi V_A + (1 - \Pi) V_S \geq \Pi V_M + (1 - \Pi) V_B \qquad (10)$$

If all four constraints are satisfied, neither agent will have an incentive to depart from a mutual truth-telling equilibrium.[11]

Having specified the truth-telling constraints, we now derive the resource constraints. In each of the four equilibrium states, the sum of firm's and worker's consumption cannot exceed the value of total output at full employment. Letting $U^{-1}(\cdot)$ and $V^{-1}(\cdot)$ denote the inverses of firm's and worker's utility functions, so that $U^{-1}(U_A)$ denotes the level of consumption which yields the firm utility U_A, etc., the four resource constraints are:

$$H \geq U^{-1}(U_A) + V^{-1}(V_A) \qquad (11)$$

$$L \geq U^{-1}(U_B) + V^{-1}(V_B) \qquad (12)$$

$$H > U^{-1}(U_M) + V^{-1}(V_M) \qquad (13)$$

$$L > U^{-1}(U_S) + V^{-1}(V_S) \qquad (14)$$

The problem is to maximize the sum of firm's and worker's expected utilities over the four equilibrium states, subject to the four truth-telling constraints and the four resource constraints. The ultimate choice variables in this problem are the contractual levels of wages and employment in each announced state, but it is more convenient to consider as choice variables the eight equilibrium contractual levels of utility, U_A, U_B, U_M, U_S, V_A, V_B, V_M, and V_S.[12]

$$\text{Max } L = \frac{\Pi}{2}(U_A + U_B + V_A + V_B) + \frac{(1 - \Pi)}{2}(U_M + U_S + V_M + V_S) \qquad (15)$$

subject to:

$$\Pi U_A + (1 - \Pi) U_M > \Pi \hat{U}_S + (1 - \Pi) \hat{U}_B, \lambda_1 \qquad (7)$$

$$\Pi \hat{U}_B + (1 - \Pi) \hat{U}_S > \Pi \hat{U}_M + (1 - \Pi) \hat{U}_A, \lambda_2 \qquad (8)$$

$$\Pi V_B + (1 - \Pi) V_M > \Pi V_S + (1 - \Pi) V_A, \lambda_3 \qquad (9)$$

$$\Pi V_A + (1 - \Pi) V_S > \Pi V_M + (1 - \Pi) V_B, \lambda_4 \qquad (10)$$

$$H > U^{-1}(U_A) + V^{-1}(V_A), \lambda_5 \qquad (11)$$

$$L > U^{-1}(U_B) + V^{-1}(V_B), \ \lambda_6 \tag{12}$$

$$H > U^{-1}(U_M) + V^{-1}(V_M), \ \lambda_7 \tag{13}$$

$$L > U^{-1}(U_S) + V^{-1}(V_S), \ \lambda_8 \tag{14}$$

The above maximization problem contains eight choice variables, three parameters (Π, H, and L), and four additional variables, $\hat{U}_A$, $\hat{U}_B$, $\hat{U}_M$, and $\hat{U}_S$. The values of these four additional variables are uniquely determined by the values of the choice variables and parameters. $\hat{U}_A$, for instance, is a function of U_A, V_A, H, and L. When maximizing with respect to U_A, therefore, we need to find the derivative of $\hat{U}_A$, with respect to U_A, holding V_A constant, etc.

$$U_A = U(N_A(H - W_A)) \tag{16}$$

$$V_A = V(N_A W_A) \tag{17}$$

$$\hat{U}_A = U(N_A(L - W_A)) \tag{18}$$

Totally differentiating (16), (17), and (18) we get:

$$dU_A = U'_A \ (dN_A(H - W_A) - dW_A N_A) \tag{19}$$

$$dV_A = V'_A \ (dN_A W_A + dW_A N_A) \tag{20}$$

$$d\hat{U}_A = \hat{U}'_A \ (dN_A(L - W_A) - dW_A N_A) \tag{21}$$

where U'_A denotes the firm's marginal utility of consumption evaluated at the contractual level of utility U_A, etc. Dividing (21) by (19) and setting (20) equal to zero, we get:

$$\frac{d\bar{U}_A}{dU_A} = \frac{\bar{U}'_A L}{U'_A H} > 0 \tag{22}$$

Dividing (21) by (20) and setting (19) equal to zero we get:

$$\frac{d\bar{U}_A}{dV_A} = \frac{\bar{U}'_A(L - H)}{V'_A H} < 0 \tag{23}$$

Similarly:

$$\frac{d\bar{U}_B}{dU_B} = \frac{\bar{U}'_B H}{U'_B L} > 0 \tag{24}$$

$$\frac{d\bar{U}_B}{dV_B} = \frac{\bar{U}'_B(H - L)}{V'_B L} > 0 \tag{25}$$

$$\frac{d\bar{U}_M}{dU_M} = \frac{\bar{U}'_M L}{U'_M H} > 0 \tag{26}$$

$$\frac{d\bar{U}_M}{dV_M} = \frac{\bar{U}'_M(L - H)}{V'_M H} < 0 \tag{27}$$

$$\frac{d\bar{U}_S}{dU_S} = \frac{U'_S H}{U'_S L} > 0 \tag{28}$$

$$\frac{dU_S}{dV_S} = \frac{\hat{U}'_S(H - L)}{V'_S L} > 0 \qquad (29)$$

The first order conditions for our maximization problem are:

$$\frac{dL}{dU_A} = \frac{\Pi}{2} + \lambda_1 \Pi - \lambda_2(1 - \Pi)\frac{d\hat{U}_A}{dU_A} - \frac{\lambda_5}{U'_A} = 0 \qquad (30)$$

$$\frac{dL}{dU_B} = \frac{\Pi}{2} + \lambda_1(1 - \Pi)\frac{d\hat{U}_B}{dU_B} + \lambda_2 \Pi - \frac{\lambda_6}{U'_B} = 0 \qquad (31)$$

$$\frac{dL}{dU_M} = \frac{(1 - \Pi)}{2} + \lambda_1(1 - \Pi) - \lambda_2 \Pi \frac{d\hat{U}_M}{dU_M} - \frac{\lambda_7}{U'_S} = 0 \qquad (32)$$

$$\frac{dL}{dU_S} = \frac{(1 - \Pi)}{2} - \lambda_1 \Pi \frac{d\hat{U}_S}{dU_S} + \lambda_2(1 - \Pi) - \frac{\lambda_8}{U'_S} = 0 \qquad (33)$$

$$\frac{dL}{dV_A} = \frac{\Pi}{2} - \lambda_2(1 - \Pi)\frac{d\hat{U}_A}{dV_A} - \lambda_3(1 - \Pi) + \lambda_4 \Pi - \frac{\lambda_5}{V'_A} = 0 \qquad (34)$$

$$\frac{dL}{dV_B} = \frac{\Pi}{2} - \lambda_1(1 - \Pi)\frac{d\hat{U}_B}{dV_B} + \lambda_3 \Pi - \lambda_4(1 - \Pi) - \frac{\lambda_6}{V'_B} = 0 \qquad (35)$$

$$\frac{dL}{dV_M} = \frac{(1 - \Pi)}{2} - \lambda_2 \Pi \frac{d\hat{U}_M}{dV_M} + \lambda_3(1 - \Pi) - \lambda_4 \Pi - \frac{\lambda_7}{V'_M} = 0 \qquad (36)$$

$$\frac{dL}{dV_S} = \frac{(1 - \Pi)}{2} - \lambda_1 \Pi \frac{d\hat{U}_S}{dV_S} - \lambda_3 \Pi + \lambda_4(1 - \Pi) - \frac{\lambda_8}{V'_S} = 0 \qquad (37)$$

plus the constraints (7) to (14) inclusive, incorporating the usual Kuhn-Tucker conditions.

It is technically difficult to solve the above problem to find the characteristics of the optimal contract for all values of Π. Instead, we will solve for the optimal contract in the limit as Π approaches unity (public information) and find how the contract varies according to small changes in Π about the value of $\Pi = 1$. Since we shall demonstrate that "strikes" occur in state S, and since the probability of state S occurring is $(1 - \Pi)/2$, what we are doing, then, is to find the characteristics of strikes, and of the contract which implies strikes, in the limit as the frequency of strikes approaches zero.

Setting $\Pi = 1$ in the first order conditions gives us:

$$\frac{1}{2} + \lambda_1 - \frac{\lambda_5}{U'_A} = 0 \qquad (30a)$$

$$\frac{1}{2} + \lambda_2 - \frac{\lambda_6}{U'_B} = 0 \qquad (31a)$$

$$-\lambda_2 \frac{d\hat{U}_M}{dU_M} - \frac{\lambda_7}{U'_M} = 0 \qquad (32a)$$

$$-\lambda_1 \frac{d\hat{U}_S}{dU_S} - \frac{\lambda_8}{U'_S} = 0 \qquad (33a)$$

$$\frac{1}{2} + \lambda_4 - \frac{\lambda_5}{V'_A} = 0 \qquad (34a)$$

$$\frac{1}{2} + \lambda_3 - \frac{\lambda_6}{V_B'} = 0 \qquad (35a)$$

$$-\lambda_2 \frac{d\bar{U}_M}{dV_M} - \lambda_4 - \frac{\lambda_7}{V_M'} = 0 \qquad (36a)$$

$$-\lambda_1' \frac{d\bar{U}_S}{dV_S} - \lambda_3 - \frac{\lambda_8}{V_S'} = 0 \qquad (37a)$$

$$U_A > \bar{U}_S \qquad (7a)$$

$$U_B > \bar{U}_M \qquad (8a)$$

$$V_B > V_S \qquad (9a)$$

$$V_A > V_M \qquad (10a)$$

plus the resource constraints (11) to (14) inclusive.

Proposition 1: $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8 \geq 0$.

Proof: Since all the constraints are inequalities, the lagrange multipliers cannot be negative.

Proposition 2: $\lambda_5 > 0$, $N_A = 1$.

Proof: From (30a), since $U_A' > 0$ and $\lambda_1 > 0$. Since $\lambda_5 > 0$, the constraint (11) holds as an equality.

Proposition 3: $\lambda_6 > 0$, $N_B = 1$.

Proof: From (31a), since $U'_A > 0$ and $\lambda_2 > 0$. Since $\lambda_5 > 0$, the constraint (12) holds as an equality.

Discussion: Propositions 2 and 3 imply that the resource constraints (11) and (12) are binding in states A and B, implying full employment in those states ($N_A = \bar{1}$, $N_B = 1$).

Proposition 4: $\lambda_1 = \lambda_8 = 0$

Proof: Consider (33a). Since $\lambda_1 > 0$ and $\lambda_8 > 0$, from proposition 1, $d\hat{U}_S/dU_S > 0$ from (28), and $U'_S > 0$, then $\lambda_1 = 0$ and $\lambda_8 = 0$ are the only values which satisfy (33a).

Proposition 5: $\lambda_2 = \lambda_7 = 0$.

Proof: Consider (32a). From proposition 1, $\lambda_2 > 0$ and $\lambda_7 > 0$. $d\hat{U}_M/dU_M > 0$ from (26), and $U'_M > 0$.

Proposition 6: $\lambda_3 = 0$.

Proof: Consider (37a) and note that $\lambda_1 = \lambda_8 = 0$ from proposition 4.

Proposition 7: $\lambda_4 = 0$.

Proof: Consider (36a) and note that $\lambda_2 = \lambda_7 = 0$ from proposition 5.

Discussion: Propositions 4, 5, 6 and 7 state that, in the limit as Π approaches one, the lagrange multipliers λ_1, λ_2, λ_3, λ_4, λ_7, and λ_8 approach zero. The intuitive meaning is that as information approaches full publicness, it costs nothing to impose truth-telling constraints, and it costs nothing to impose resource constraints in states M and S when the probability of those states' occurring vanishes.

Proposition 8: $W_A = H/2$.

Proof: From (30a) and (34a), since $\lambda_1 = \lambda_4 = 0$ from propositions 4 and 7, it follows that $U'_A = V'_A$. Since firm and worker have identical risk-averse utility functions, they must therefore enjoy equal consumption levels in state A. Since firm's consumption $N_A(H - W_A)$ equals worker's consumption $N_A W_A$, it follows that $W_A = H/2$.

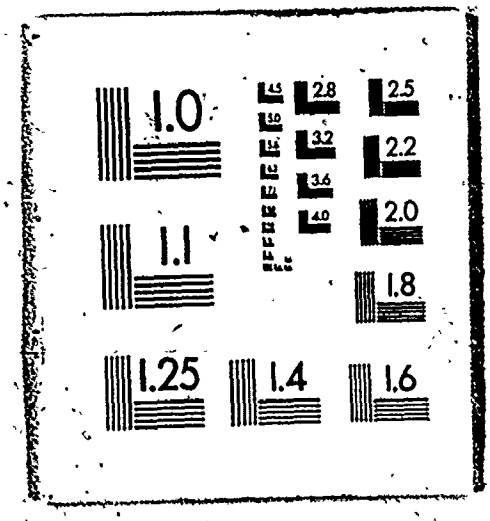
Proposition 9: $W_B = L/2$.

Proof: From (31a) and (35a), since $\lambda_2 = \lambda_3 = 0$ from propositions 5 and 6, it follows that $U'_B = V'_B$. Since firm and worker have identical risk-averse utility functions, their respective levels of consumption in state B, $N_B(L_B - W_B)$ and $N_B W_B$, must be equal. Hence $W_B = L/2$.

Discussion: Propositions 8 and 9 state that in the limit as Π approaches one, the contract exhibits perfect risk-sharing in states A and B, the only two states with non-vanishing probabilities of occurrence. Together with propositions 2 and 3, implying full employment in states A and B, this serves as a check on our results by demonstrating that the outcomes of the private information contract, in the limit as information approaches full publicness, conform to the outcomes of the public information contract. We are more interested, however, in the contractual provisions for state S, the probability of which (and of state M) vanishes in the limit.

Proposition 10: $N_S < (1 + L/H)/2 < 1$.

Proof: From (7a) and proposition 2 ($N_A = 1$), we deduce that $(H - W_A) > N_S(H - W_S)$. From (9a) and proposition 3 ($N_B = 1$), we deduce that $W_B > N_S W_S$. These together imply that $N_S < 1 - (W_A - W_B)/H$. Substituting for



W_A and W_B from propositions 8 and 9 we get: $N_S < (1 + L/H)/2$. Since, by assumption, $L < H$, this implies that $N_S < 1$.

Discussion: Proposition 10 is a central result of the model. It states that, in the limit as Π approaches one, the resource constraint in state S, (14), will hold as a strict inequality; there will be less than full employment in that state. This underemployment is ex post inefficient, but efficient ex ante, there being no way to preserve truth-telling in a risk-sharing (variable wage) contract without the threat of a destruction of rents should the firm announce "L" (implying a low wage) and the worker announce "H," (implying a high wage).

Thus far, we have solved for wages and employment in states A and B, and have found an upper bound for employment in state S, these results being valid in the limit as Π approaches one. We need to solve for wages and employment in state M, for wages in state S, and to check whether employment in state S will equal its upper bound as given by proposition 10.

We are unable to derive these results from the set of equations found by substituting $\Pi = 1$ into the first order conditions. In terms of the algebra of the problem, the reason for this is that the attempt to do so would involve evaluating expressions involving zero divided by zero, etc. In terms of the economics of the problem, at the limit when $\Pi = 1$ the states M and S never occur, and so the levels of employment and wages in those states have no direct effect on the parties' expected utilities.

To circumvent this problem we differentiate the first order conditions with respect to Π, and evaluate the resulting set of equations at

$\Pi = 1$. This procedure also enables us to derive predictions concerning the effects on the contract of small changes in Π about $\Pi = 1$.

Differentiating the first order conditions (equations (30) to (37) and (7) to (14)) with respect to Π, and substituting the results from the first order conditions evaluated at $\Pi = 1$ (equations (30a) to (37a), and (7a) to (10a), plus propositions 1 to 10) we get:

$$\frac{1}{2} + \frac{d\lambda_1}{d\Pi} - \frac{d\lambda_5}{d\Pi U'_A} + \frac{dU'_A}{d\Pi 2U'_A} = 0 \qquad (30b)$$

$$\frac{1}{2} + \frac{d\lambda_2}{d\Pi} - \frac{d\lambda_6}{d\Pi U'_B} + \frac{dU'_B}{d\Pi 2U'_B} = 0 \qquad (31b)$$

$$-\frac{1}{2} - \frac{d\lambda_2}{d\Pi}\frac{d\hat{U}_M}{dU_M} - \frac{d\lambda_7}{d\Pi U_M} = 0 \qquad (32b)$$

$$-\frac{1}{2} - \frac{d\lambda_1}{d\Pi}\frac{d\hat{U}_S}{dU_S} - \frac{d\lambda_8}{d\Pi U'_S} = 0 \qquad (33b)$$

$$\frac{1}{2} + \frac{d\lambda_4}{d\Pi} - \frac{d\lambda_5}{d\Pi V'_A} + \frac{dV'_A}{d\Pi 2V_A} = 0 \qquad (34b)$$

$$\frac{1}{2} + \frac{d\lambda_3}{d\Pi} - \frac{d\lambda_6}{d\Pi V'_B} + \frac{dV'_B}{d\Pi 2V'_B} = 0 \qquad (35b)$$

$$-\frac{1}{2} - \frac{d\lambda_2}{d\Pi}\frac{d\hat{U}_M}{dV_M} - \frac{d\lambda_4}{d\Pi} - \frac{d\lambda_7}{d\Pi V'_M} = 0 \qquad (36b)$$

$$-\frac{1}{2} - \frac{d\lambda_1}{d\Pi}\frac{d\bar{U}_S}{dV_S} - \frac{d\lambda_3}{d\Pi} - \frac{d\lambda_8}{d\Pi V'_S} = 0 \qquad (37b)$$

$$U_A + \frac{dU_A}{d\Pi} - U_M - \hat{U}_S - \frac{d\bar{U}_S}{d\Pi} + \hat{U}_B = 0 \quad \text{or} \quad \frac{d\lambda_1}{d\Pi} = 0 \qquad (7b)$$

$$U_B + \frac{dU_B}{d\Pi} - U_S - \hat{U}_M - \frac{d\bar{U}_M}{d\Pi} + \hat{U}_A = 0 \quad \text{or} \quad \frac{d\lambda_2}{d\Pi} = 0 \qquad (8b)$$

$$V_B + \frac{dV_B}{d\Pi} - V_M - V_S - \frac{dV_S}{d\Pi} + V_A = 0 \quad \text{or} \quad \frac{d\lambda_3}{d\Pi} = 0 \qquad (9b)$$

$$V_A + \frac{dV_A}{d\Pi} - V_S - V_M - \frac{dV_M}{d\Pi} + V_B = 0 \quad \text{or} \quad \frac{d\lambda_4}{d\Pi} = 0 \qquad (10b)$$

$$\frac{dU_A}{d\Pi U'_A} + \frac{dV_A}{d\Pi V'_A} = 0 \qquad (11b)$$

$$\frac{dU_B}{d\Pi U'_B} + \frac{dV_B}{d\Pi V'_B} = 0 \qquad (12b)$$

$$\frac{dU_M}{d\Pi U'_M} + \frac{dV_M}{d\Pi V'_M} = 0 \quad \text{or} \quad \frac{d\lambda_3}{d\Pi} = 0 \qquad (13b)$$

$$\frac{dU_S}{d\Pi U'_S} + \frac{dV_S}{d\Pi V'_S} = 0 \quad \text{or} \quad \frac{d\lambda_8}{d\Pi} = 0 \qquad (14b)$$

Discussion: If, for some lagrange multiplier λ, it is the case that $d\lambda/d\Pi < 0$, and that $\lambda = 0$ at $\Pi = 1$, we know that the constraint asso-

ciated with that multiplier will be binding for values of Π slightly less than one so that he constraint will hold as an equality in the limit, hence the form of (7b) to (14b), except for (11b) and (12b). For (11b) and (12b), since propositions 2 and 3 show that λ_5 and λ_6 are strictly positive, we already know that these constraints hold as equalities in the limit as Π approaches one.

Proposition 11: $d\lambda_8/d\Pi = 0$.

Proof: From proposition 10 we know that the resource constraint in state S, (14), holds as a strict inequality when $\Pi = 1$. Given continuity, there must therefore be a neighbourhood about $\Pi = 1$ such that $\lambda_8 = 0$ within that neighbourhood.

Proposition 12: $d\lambda_1/d\Pi < 0$.

Proof: Consider equation (33b), noting that $d\lambda_8/d\Pi = 0$ from proposition 11, and that $dO_S/dU_S > 0$ from equation (28).

Proposition 13: $O_S = U_A$.

Proof: Propositions 4 and 12 tell us that λ_1 approaches zero from positive values as Π approaches one. Therefore, constraint (7) is binding in the neighbourhood of $\Pi = 1$, so that (7a) holds as an equality in the limit.

Proposition 14: Either (a) $d\lambda_3/d\Pi = 0$ and $V_S'/U_S' = (H - L)/H$ or (b) $d\lambda_3/d\Pi < 0$ and $V_S = V_B$.

Proof: From proposition 1, $\lambda_3 > 0$, and from proposition 6, $\lambda_3 = 0$ when $\Pi = 1$, therefore, at $\Pi = 1$, either $d\lambda_3/d\Pi = 0$ or $d\lambda_3/d\Pi < 0$.

(a) Noting that $d\lambda_8/d\Pi = 0$ from proposition 11, solving for $d\lambda_1/d\Pi$ from (33b) and substituting into (37b) we get:

$$\frac{d\lambda_3}{d\Pi} = -\frac{1}{2} + \frac{1}{2}\frac{dU_S}{d\hat{U}_S}\frac{d\hat{U}_S}{dV_S} \tag{38}$$

Substituting for $d\hat{U}_S/dU_S$ from (28) and for $d\hat{U}_S/dV_S$ from (29) we get:

$$\frac{d\lambda_3}{d\Pi} = -\frac{1}{2} + \frac{1}{2}\frac{U'_S}{V'_S}\frac{(H - L)}{H} \tag{39}$$

If $d\lambda_3/d\Pi = 0$, then (39) gives us: $V'_S/U'_S = (H - L)/H$.

(b) If $d\lambda_3/d\Pi < 0$, then, since $\lambda_3 = 0$ at $\Pi = 1$, from proposition 6, we know that λ_3 approaches zero from positive values as Π approaches one. Therefore, constraint (9) is binding in the neighbourhood of $\Pi = 1$, so (9a) holds as an equality in the limit, implying $V_S = V_B$.

Discussion: Propositions 12 and 13 tell us that the firm's truth-telling constraint (7) will be binding for values of Π near one. We are unable to prove that the corresponding constraint (9) on the worker will be binding. The solution for V_S implicit in proposition 14a may or may not violate constraint (9a), depending on the form of the utility function and the relative magnitudes of the parameters H and L. The smaller the difference between H and L, and the less the degree of risk aversion, the larger will be the value of V_S according to proposition 14a, and so the more likely would the constraint (9a) be binding. Furthermore, as the value of $(H - L)/H$ becomes arbitrarily small, the value of V_S, for any

utility function with positive and diminishing marginal utility, becomes arbitrarily large, hence we will assume from now on that proposition 14b, as opposed to 14a, is applicable; that constraint (9a) holds as an equality.

Proposition 15: $N_S = (1 + L/H)/2$.

Proof: Noting propositions 13 and 14b, the proof proceeds as for proposition 10, with equalities replacing inequalities.

Proposition 16: $W_S = LH/(L + H)$.

Proof: Propositions 9 and 14b imply that $N_S W_S = L/2$. Now substitute for N_S from proposition 15.

Discussion: Propositions 15 and 16 determine the contractual levels of employment and wages, as Π approaches one in the limit, in the "strike" state S. They are determined by the truth-telling constraints (7) and (9), and by the contractual levels of wages in states A and B. Note that $N_S < 1$ and $W_A > W_S > W_B$.

Proposition 17: $d\lambda_2/d\Pi = 0$, $d\lambda_4/d\Pi = 0$, $d\lambda_7/d\Pi = -\frac{1}{2}$, $N_M = 1$, and $W_M = H/2$.

Proof: Suppose that $d\lambda_2/d\Pi = 0$ and that $d\lambda_4/d\Pi = 0$. From (32b) and (36b) we deduce that $U'_M = V'_M$ (implying $U_M = V_M$) and that $d\lambda_7/d\Pi = -\frac{1}{2}$. The latter, since λ_7 approaches zero at $\Pi = 1$, from proposition 5, implies that the resource constraint (13) holds as an equality, implying that $N_M = 1$. If $U'_M = V'_M$, since firm and worker have identical risk averse utility functions, their respective levels of consumption in state

M, $N_M(H - W_M)$ and $N_M W_M$, must be equal, which implies that $W_M = H/2$. From propositions 2 and 8, this in turn implies that $U_M = U_A$, and $V_M = V_A$, which satisfies constraint (16a). From propositions 3 and 9 we know that $U_B = U(L/2)$, which exceeds $U_M = U(L - H/2)$, and so satisfies the constraint (14a). Since the supposition that $d\lambda_2/d\Pi = 0$ and $d\lambda_4/d\Pi = 0$ permitted us to derive values for U_M and V_M which do not violate the corresponding constraints (14a) and (16a), we have validated our supposition.

Discussion: Proposition 17 gives us the contractual levels of employment and wages in state M in the limit as Π approaches one. Note that they involve full employment and efficient risk-sharing, as do states A and B. We have now solved for wages and employment in all four announced states.

Proposition 18: $dU_A/d\Pi < 0$, $dV_A/d\Pi > 0$.

Proof: Since firm and worker have identical risk averse utility functions, and since $U_A = V_A$ from propositions 2 and 8, equation (11b) implies:[13]

$$\frac{dU'_A}{d\Pi U'_A} = -\frac{dV'_A}{d\Pi V'_A} \tag{40}$$

Subtracting (34b) from (30b), noting that $d\lambda_4/d\Pi = 0$ from proposition 17, and using (40), we get:

$$\frac{dU'_A}{d\Pi} = -\frac{dV_A}{d\Pi} = -U'_A \frac{d\lambda_1}{d\Pi} \tag{41}$$

Since $d\lambda_1/d\Pi < 0$ from proposition 12, and since $dU_A/d\Pi$ and $dV_A/d\Pi$ are respectively opposite in sign from $dU'_A/d\Pi$ and $dV'_A/d\Pi$, we deduce that $dU_A/d\Pi < 0$ and $dV_A/d\Pi > 0$.

Proposition 19: $dU_B/d\Pi > 0$, $dV_B/d\Pi < 0$.

Proof: Since firm and worker have identical risk averse utility functions, and since $U_B = V_B$ from propositions 3 and 9, equation (12b) implies that:

$$\frac{dU'_B}{d\Pi U'_B} = -\frac{dV'_B}{d\Pi V'_B} \tag{42}$$

Subtracting (35b) from (31b) and using (42), noting that $d\lambda_2/d\Pi = 0$ from proposition 17, we get:

$$\frac{dU'_B}{d\Pi} = -\frac{dV_B}{d\Pi} = \hat{U}_B \frac{d\lambda_3}{d\Pi} \tag{43}$$

If the proposition (14b) holds, so that $d\lambda_3/d\Pi < 0$, and since $dU_B/d\Pi$ and $dV_B/d\Pi$ are respectively opposite in sign from $dU'_B/d\Pi$ and $dV'_B/d\Pi$, we deduce that $dU_B/d\Pi > 0$ and $dV_B/d\Pi < 0$.

Discussion: In the limit as Π approaches one (public information) the contract approaches full risk-sharing between firm and worker. Propositions 18 and 19 tell us that as Π falls below one (partially private information) we get an increase in U_A (fall in V_A) and a fall in U_B (increase in V_B), meaning that the firm takes on a progressively greater share of the risk. The reason for this result is that shifting more of

the risk from firm to worker implies having greater variation of the wage across announced states A and B, which increases the benefits from making a false announcement, and so requires an increase in the costliness of strikes in order to preserve the incentive for truth-telling. When $\Pi = 1$, the costliness of strikes does not matter, for strikes (state S) never occur in equilibrium, and so full risk-sharing is efficient. As Π falls below one, strikes do occur in equilibrium, so the optimum contract trades-off some risk-sharing in order to reduce the costliness of strikes, and hence produces imperfect risk-sharing.

IIc. The Concession Curves

In section IIb we formalized the problem the firm and worker face in choosing an efficient contract under private information. We solved for the efficient contract in the limit as the parameter Π approaches one (as information approaches full publicness). We found in one state, state S, where the firm observes a low price and the worker observes a high price, that the contractual level of employment would be less than the ex post efficient level of employment. The reason for this underemployment is to preserve the parties' incentives for truth-telling - hence to preserve a contract which shares risk by indexing the wage to the price of the firm's output. Are we justified, however, in interpreting the under-employment in state S as constituting a strike?

There are two features of actual strikes which any theory which claims to be a theory of strikes, as opposed to (say) layoffs, must capture. The first is the heavily normative content of the parties' pronouncements before and during the strike. Each side claims that its proposed wage is, in some sense, just, and that the other's proposal is

unjust. Each claims that it deserves a larger share of the joint rents than the other is offering. Some might argue that these claims must not be taken at face value, that they are mere window dressing over what is, in reality, a simple exercise of bargaining power. If strikes really were just an exercise in bargaining power, however, why should the parties pretend it to be anything other? Who are they trying to fool, and why? The theory proposed in this paper can give credence to the normative content of the parties' pronouncements. Suppose that state S occurs, that the worker observes a high price and the firm observes a low price. As the parameter Π approaches one, the worker, on observing P_W = H, believes with a probability approaching certainty that the true state is A, and that the contractual, or "just," wage is W_A. The firm on observing P_F = L, believes with a probability approaching certainty that the true state is B, and that the "just" wage is W_B. If one holds, or assumes firm and worker to hold, a contractarian theory of justice, then the normative content of their pronouncements during a strike becomes fully understandable.

The second feature of actual strikes is that both parties put forward a wage-offer or wage-demand on the eve of the strike, and generally concede during the progress of the strike, which ends when the two proposed wages coincide. The existence of these concession curves, and their shapes and positions, is predicted by the theory of strikes put forward in this paper. The firm cannot rely on the worker to impose a truth-telling constraint on himself; the firm must be able to do this unilaterally. In the limit as Π approaches one, the worker's truth-telling constraint becomes:

$$V_B > V_S \qquad (9a)$$

which implies:

$$N_B W_B > N_S W_S. \tag{44}$$

Since $N_B = 1$, defining the strike length $S \equiv 1 - N$, we can rearrange (44) as:

$$W_S < W_B/(1 - S) \tag{45}$$

To impose truth-telling on the worker, the firm, on observing $P_F = L$, must ensure that the worker's rents do not exceed W_B (the rents the worker would get if he announces $P_W = L$). The firm therefore makes an initial offer of W_B, and then concedes slowly during the progress of the strike in accordance with (45).

Similarly, the worker unilaterally imposes the firm's truth-telling constraint:

$$U_A > U_S \tag{7a}$$

which implies:

$$N_A(H - W_A) > N_S(H - W_S) \tag{46}$$

Since $N_A = 1$, rearranging gives:

$$W_S > (W_A - HS)/(1 - S) \tag{47}$$

To impose truth-telling on the firm, the worker, on observing $P_W = H$,

must ensure that the firm's rents do not exceed $(H - W_A)$, (the rents the firm would get if it announces $P_F = H$). The worker therefore makes an initial demand of W_A, and concedes during the strike in accordance with (47).

Treating (45) and (47) as equalities,[14] the former gives the firm's wage-offer curve and the latter the worker's wage-demand curve. Solving the two curves simultaneously gives us the strike length and settlement wage:

$$S = \frac{W_A - W_B}{H} \tag{48}$$

$$W_S = W_B H/(H - W_A + W_B) \tag{49}$$

Which, substituting $W_A = H/2$ and $W_B = L/2$, agree with the solutions for N_S and W_S given by propositions 15 and 16.

The two concession curves are depicted in Figure Two. The firm's wage-offer curve (45) is a rectangular hyperbola about the point ($S = 1$, $W = 0$), thus exhibiting constant rents to the worker at all points along it. The worker's wage-demand curve (47) is a rectangular hyperbola about the point ($S = 1$, $W = H$), thus exhibiting constant rents to the firm if the price of the firm's output were indeed H.

It is important to distinguish the concession curves predicted by this theory from those found in Hicks (1964, p. 143). The interpretation and equations defining the curves are quite different. Hicks' "employer's concession curve" for instance, is defined as the set of points $\{W(S), S\}$ such that the firm's rents if it pays $W(S)$ with no strike equal its rents if it pays $W(0)$ after a strike of length S. My

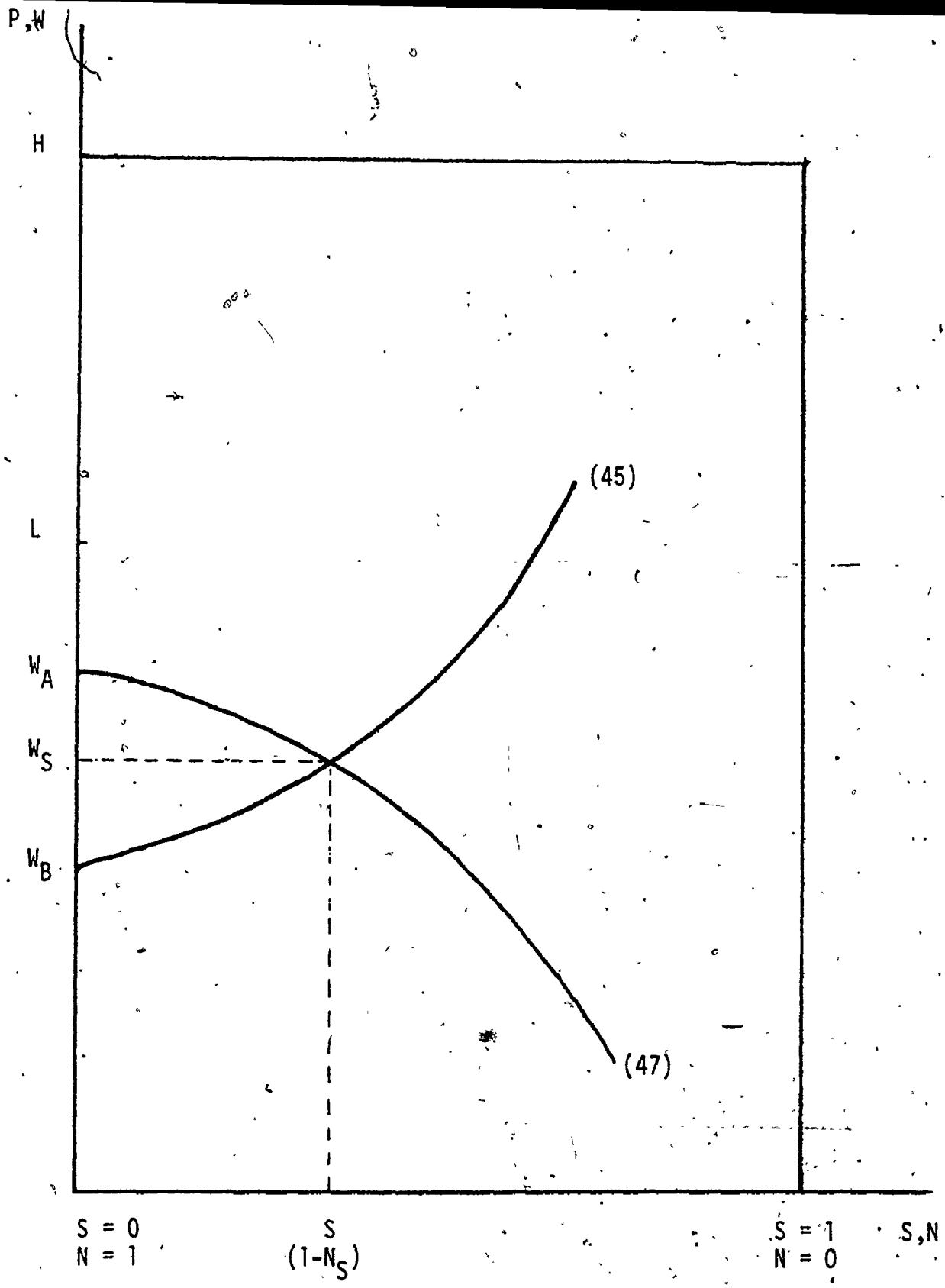


Figure Two

own firm's wage-offer curve is also upward sloping, but is defined as the set of points {W(S), S} such that the worker's rents if he accepts W(S) after a strike of length S equal his rents if he accepts W(0) with no strike." Given the assumptions of constant returns and no disutility of labour, the Hicksian concession "curves" (it can be verified) are straight lines. My own concession curves are indeed curves, corresponding to the concave yield curves discovered empirically by Comay, Melnik and Subotnik (1974).

III. Conclusion

Strikes involve the conspicuous failure of agents to trade despite the apparent existence of potential gains from trade. Presumably there exists some wage such that both sides could gain by trading at that wage relative to what each can rationally expect to gain by continuing the strike. At least one side is apparently irrational in their refusal to reach such an agreement. Why then do strikes exist?

As with other cases of apparently irrational behaviour, this paradox can be resolved if we see the apparently irrational action as part of a rational rule of action. It seems irrational to imprison a convicted criminal, since this imposes costs both on the prisoner and on the rest of society, but it is rational to follow a rule of punishing convicted criminals, since the expectations created by adherence to such a rule have the benefit of detering crime.[15] Similarly, the worker (or firm) in refusing to trade at a wage below (above) a certain level, appears irrational, but it is rational for him to follow a rule, requiring him to act thus, since the expectations created by his adherence to such a rule have the benefit of enforcing a contract which could not otherwise be maintained. Though this feature has not been stressed in the above-formulated model, the only reason a worker (or firm) would incur the cost of following a strategy of slow concession during a strike (as opposed to immediate concession) is that by doing so he maintains his reputation for doing so - he promotes the expectation in this and other future potential trading partners that he will act likewise in future - thereby preserving his ability to enter confidently into future trades governed by that same sort of contract.

The essential feature of the theory of strikes presented here is that two or more parties wish to enter a contractual arrangement where they cannot tell with certainty ex ante whether or not a breach of contract has occurred. In order to deter breaches, each must act as if a perceived breach were in fact an actual breach, and take steps to ensure that the breach is not rewarded, by constraining the perceived violator's rents not to exceed the rents under no perceived breach. Since the attempt to punish a perceived breach must itself be treated as a perceived breach by the other, each gets the rents the other perceives him to deserve, and the disputed rents are destroyed by the strike.

The model presented in this paper is intended merely to provide a formal example of this theory of strikes. Many of the model's features are not essential. We could allow the variable privately observed by the worker to matter intrinsically - not merely as a provider of information about the variable privately observed by the firm. The optimal contract might specify a wage or other variable contingent on variables other than the price of the firm's output (e.g., productivity, the price of the workers' consumption, other wages, etc.); provided only that there is some asymmetry of information sufficient to ensure that both do not always agree on the public information contractual wage. The contract could be motivated by considerations other than risk - aversion. Furthermore, the theory need not be confined to the analysis of conflict in the market for labour, but could be extended to explain other examples of costly conflict.

Footnotes

* This paper has been through several drafts since the first version in 1980. I would like to thank Richard Brecher, Roger Farmer, Joel Fried, David Laidler, Tim Lane, Glenn MacDonald, Felice Martinello, Chris Robinson, Tom Rymes, and Ron Wintrobe for their assistance and comments on earlier drafts. All responsibility for errors and opinions is my own.

1. An alternative answer is provided by Stigler (1970); too high a penalty for trivial crimes would cause criminals to switch to more serious crimes.

2. Suppose that capital punishment were imposed for breaking the speed limit? Drivers would take unreasonable precautions such as either driving very slowly or else investing in supremely accurate and reliable (and hence costly) speedometers.

3. Or else invest inefficiently large amounts in gaining information on the price level.

4. The labour market is thus ex ante competitive, but exhibits bilateral monopoly between each pair ex post. The isolatedness of the islands, which ensures the latter, is intended as a parable for any fixed costs in setting up a trading relation between particular partners, e.g., hiring costs, or firm-specific training.

5. The assumption of constant returns is for convenience, allowing us to talk of (normalized) labour and output indiscriminately. Under non-constant returns, the model remains valid provided N is interpreted as output, with corresponding levels of employment given by the production function, and W interpreted as a piece-rate rather than hourly wage.

6. P_F is thus a relative price, and W a real wage, with the consumption good as numeraire.

7. The labour market is thus like the Rawlsian "state of nature," except that agents maximize expected utility rather than maximizing minimum utility over roles; hence the utilitarian objective function of the contract.

8. We need not consider values for $\Pi < 1/2$, since e.g., $\Pi = 0$ implies a perfect negative correlation between P_F and P_W, and thus is equivalent to public information of $\Pi = 1$.

9. Thus U_A, for example, designates $U(N_A(L - W_A))$, the firm's utility if the firm and worker both announce a high price, but the firm observes a low price.

10. The "leaders" constraints are unchanged, and the "followers" constraints are found by taking his simultaneous announcement constraints and imposing them both when $\Pi = 0$ and when $\Pi = 1$, thus giving the follower four truth-telling constraints. The space permitted by the six sequential constraints is a subset of the space

permitted by the four simultaneous constraints, since the latter are a convex combination of the former.

11. Notice, however, that each tells the truth (lies) if and only if he expects the other to tell the truth (lie). Thus mutual lying is also an equilibrium. However, in this case we simply switch the labels of A and B and of M and S. It does not matter if we use the word "high" to mean low and vice versa, providing we stick to some convention. Compare this game to the game of driving on the right or left side of the road.

12. Since U_A and V_A are each functions of N_A and W_A, we can deduce the latter pair from the former pair, etc.

13. Strictly speaking, we need to assume that U and V are twice continuously differentiable for this result.

14. As propositions 13 and 14b permit us to do.

15. At root, the paradox is that of the time-inconsistency of optimal plans under rational expectations, and the advantages of commitment to a rule as opposed to discretion, as explained by Kydland and Prescott (1977). I have explored the rationality of following a rule in Rowe (1982).

References

Ashenfelter, O. and G.E. Johnson (1960), "Bargaining Theory, Trade Unions, and Industrial Strike Activity," American Economic Review, 59 (March 1969).

Comay, Y., A. Melnik, and A. Subotnik (1974), "Bargaining, Yield Curves, and Wage Settlement: An Empirical Analysis," Journal of Political Economy, 82 (March 1974).

Crawford, V.P. (1982), "A Theory of Disagreement in Bargaining," Econometrica, 50 (May 1982).

Cross, J.G. (1965), "A Theory of the Bargaining Process," American Economic Review, 55 (March 1965).

Grossman, S.J. and O.D. Hart (1981), "Implicit Contracts, Moral Hazard, and Unemployment," American Economic Review, 71 (May 1981).

Hayes, B. (1984), "Unions and Strikes with Asymmetric Information," Journal of Labour Economics, 2 (January 1984).

Hicks, J.R. (1964), The Theory of Wages, 2nd. ed., London: MacMillan.

Kydland, F.E., and E.C. Prescott (1977), "Rules rather than Discretion: The Inconsistency of Optimal Plans," Journal of Political Economy, 85 (May 1977).

Rowe, N. (1982), "The Rationality of Promises," Unpublished.

Stigler, G.J. (1970), "The Optimum Enforcement of Laws," Journal of Political Economy, 78 (May/June 1970).

END

12-11-85

FIN