IMPLEMENTATION AND RENEGOTIATION*

by

Eric Maskin
Harvard University

and

John Moore
London School of Economics and University of St. Andrews

Contents:

Abstract
1. Introduction
2. Renegotiation
3. Lotteries
4. The Case of Two Agents
5. More than Two Agents
References

Discussion Paper
No. TE/98/366
December 1998

The Suntory Centre
Suntory and Toyota International Centres
for Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street, London WC2A 2AE
Tel.: 020 7955 6698

---

* This work was supported by the National Science Foundation and the Leverhulme Trust. The paper has existed in draft form since September 1987. The original version benefited from comments by Dilip Abreu and Patrick Rey.

Abstract

The paper characterises the choice rules that can be implemented when agents are unable to commit themselves not to renegotiate the mechanism.

Keywords: social choice rule; implementation; renegotiation.

JEL No.: D78

Contact address: Professor Eric Maskin, School of Social Science, Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540. Email: maskin@ias.edu

1. Introduction

Implementation theory attempts to answer the question: When is it possible to design a game form (also called a mechanism or outcome function) whose equilibrium outcomes are assured of being optimal with respect to some given criterion of social welfare (called a social choice rule)?[1] Formally, consider a "society" of n agents, $1,\ldots,n$, and let A be a finite set of social alternatives (an "alternative" might, for example, be an allocation of goods in economic contexts or a public decision in political contexts).[2] Let Θ be a set of possible states of the world. A state θ includes all relevant information about preferences, endowments, and so on. Hence, let $R_i(\theta)$ denote agent i's preferences over A in state θ, for $i = 1,\ldots,n$.[3]

---

[1] For recent surveys of implementation theory, see Corchon [1996], Moore [1992], Palfrey [1998], and Chapter 10 in Osborne and Rubinstein [1994].

[2] We assume A is finite so as to ensure that most-preferred alternatives always exist. The assumption is inessential, and is made only to simplify the presentation.

[3] We shall use the notation "$aR_i(\theta)b$" to mean that alternative a is ranked at least as high as alternative b in preference ordering $R_i(\theta)$ (i.e., a is weakly preferred to b under $R_i(\theta)$). The notation "$aP(R_i(\theta))b$" means that a is strictly preferred to b under $R_i(\theta)$. And "$aI(R_i(\theta))b$" means that a and b are indifferent in preference ordering $R_i(\theta)$.

A social choice rule (SCR) f is a correspondence

$$f:\Theta \twoheadrightarrow A,$$

where $f(\theta)$ $(\subseteq A)$ is interpreted as the set of alternatives that are deemed socially optimal in state θ.[4] We call the elements of $f(\theta)$ the f-optima in state θ. The implementation problem consists of constructing a game form (which we sometimes refer to as a mechanism) g such that, for any state θ, the set of equilibrium outcomes, i.e., the set $EQ_g(\theta) = \{a|a$ is an equilibrium outcome for g in state $\theta\}$ coincides with $f(\theta)$:

$$(*) \qquad EQ_g(\theta) = f(\theta).^5$$

Implicit in this formulation is the assumption that the game form g must be designed <u>before</u> the state of the world is known, i.e., before θ is realized. Condition (*) then ensures that, whatever this state turns out to be, the outcome produced by g will be f-optimal.

---

[4] In general, the set of feasible alternatives will itself depend on θ (as when endowments or technologies differ across states; see Hurwicz, Maskin, and Postlewaite [1995]). But we shall ignore this issue here.

[5] A game form is just an ordinary game tree except that, at each terminal node, there is a physical outcome rather than a payoff vector. That is, a game form is a game in which agents' preferences over outcomes are left unspecified.

One important requirement is that the outcome of g cannot directly be a function of the state θ. (Indirectly, of course, it can depend on θ, since equilibrium strategies will depend on the state.) The rationale for this constraint is that θ entails subjective information, e.g. about preferences, and such information is not typically verifiable. (Indeed, if θ were verifiable, so that we could make the outcome of g depend directly on θ, there would exist a trivial mechanism implementing f -- at least when f is single-valued -- namely, $g(\theta) \equiv f(\theta)$.) Although θ is unverifiable, we suppose that θ, once it is realized, is common knowledge among the agents. Hence we are in the realm of "complete information" implementation, as opposed to Bayesian implementation.[6]

Of course, the set $EQ_g(\theta)$ depends on the equilibrium concept. In this paper, we will concentrate for the most part on subgame-perfect equilibrium (c.f., Moore and Repullo [1988] and Abreu and Sen [1990]), although Theorem 5 pertains to Nash equilibrium and Theorems 1 and 2 to any refinement of Nash equilibrium.

Often the mechanism g is interpreted as being imposed by a "planner" charged with maximizing social welfare. But this is not the only possible interpretation. We could, for example, conceive of the agents themselves as choosing g before the state θ is realized. In that case, if n is large, we might refer to g as a constitution. And, if n is small (e.g., n = 2), we could think of it as a contract between the agents. We shall particularly

---

[6]For a survey of the Bayesian implementation literature, see Palfrey [1992].

emphasize this last interpretation.

With few exceptions (see, for example, Aghion, Dewatripont and Rey [1994], Chung [1992], Green and Laffont [1988], Noldecke and Schmidt [1995], Hart and Moore [1988, 1999], Hermalin and Katz [1991], Maskin and Tirole [1999], Rubinstein and Wolinsky [1992], and Segal [1999]), the literature on implementation has ruled out the possibility of renegotiation. This is an issue that arises largely as an out-of-equilibrium phenomenon. When agents design a constitution or contract, they are presumably interested in ensuring Pareto optimal outcomes, and so an equilibrium outcome of the implementing mechanism will be efficient in this sense, that is, there will be no scope for renegotiation. But out of equilibrium, outcomes might be quite far from being Pareto efficient. In fact, mechanisms in the implementation literature sometimes work by assigning very bad outcomes to out-of-equilibrium strategy profiles as a way of discouraging deviations. Yet suppose that agents find themselves in an out-of-equilibrium position in which they are faced with the prospect of an inefficient outcome. Why should they put up with this when there is another possible outcome that they all prefer? In other words, why shouldn't they simply tear up their contract and renegotiate a new one in order to realize this Pareto improvement?

Unfortunately, what happens out of equilibrium can profoundly affect what outcomes can occur in equilibrium. In the absence of renegotiation, we might be able to sustain an outcome as an equilibrium by threatening agents with dire consequences should any of them deviate. But if an agent forecasts that those unfavorable consequences would ultimately be renegotiated, he might no longer have sufficient incentive to conform.

So the possibility of renegotiation should be thought of as a constraint on what outcomes can arise in equilibrium, i.e., on what SCRs can be implemented.[7] Let us consider a simple example.

Agent 1's preferences		Agent 2's preferences	
θ	ϕ	θ	ϕ
a	a	c	b
c	c	a	a
b	b	b	c

TABLE 1
A Two-Person Example

In Table 1, we have indicated the preferences of two agents over three alternatives {a,b,c} in two states, θ and ϕ. For example, in state θ, agent 2 prefers c to a and a to b. (Note that agent 1's preferences are the same in either state.) Consider the social choice rule f such that $f(\theta) = \{a\}$ and $f(\phi) = \{b\}$. If we leave aside the issue of renegotiation for a moment, there is a simple mechanism that implements f, namely, to have agent 2 choose between a and b. Observe that he should choose a in state θ (since he

---

[7] For that reason, one of us has argued that fully rational parties ought to be able to commit themselves not to renegotiate (see Maskin and Tirole (1999)). For the other's rebuttal to that position see Hart and Moore (1999). In this paper we shall assume that any such commitment is impossible.

prefers a to b in θ) and b in state ϕ (since he prefers b to a in ϕ), which is why the mechanism works. But what if agent 2 happened to choose b in state θ? Note that b is inefficient; it is Pareto-dominated by both a and c. If b were renegotiated to a, there would be no problem: in that case, whether agent 2 chose a or b in state θ, agents would end up with a, the f-optimal outcome. However, what if b were always renegotiated to c in state θ? In that case, agent 2 would deliberately choose b in state θ, anticipating that it would then be renegotiated to c. Clearly, with this latter sort of renegotiation, the simple mechanism no longer serves to implement f (indeed, it is an immediate consequence of Theorem 1 below that, if b is always renegotiated to c in state θ, no mechanism can implement f.)

This example illustrates that not only can the possibility of renegotiation constrain the set of implementable SCRs, but that the nature of the renegotiation (e.g., whether b gets renegotiated to a or c in state θ) may be crucially important. In section 2 we propose a way of formalizing renegotiation that we believe is sufficiently broad to include almost all models in the existing literature. Then, having explained the role that lotteries might play in mechanism design (section 3), we go on in sections 4 and 5 to characterize the implementable SCRs in the $n = 2$ and $n > 2$ cases respectively.

2. Renegotiation

We can think of the renegotiation process as a game. However, unlike the mechanism g -- which is expressly designed -- it is given exogenously.[8] It is determined by agents' relative bargaining strengths, their outside opportunities, and so on.

We shall model this renegotiation game as a "black box." Suppose that, in state θ, the mechanism g results in alternative a. If a is not Pareto optimal, there are gains from renegotiation. Presumably, the outcome reached in the equilibrium of the renegotiation game depends on a, since a serves as the default outcome (the "threat point") should negotiation break down. Clearly, the agents' preferences over different alternatives will also affect the negotiation, i.e., the outcome depends on the state θ.

Hence, we will suppose that the renegotiation process can be expressed as a function

$$h: A \times \Theta \longrightarrow A,$$

---

[8] In this respect we differ from such papers as Aghion, Dewatripont, and Rey (1994), who ask what happens if parties attempt to control the renegotiation process contractually, so that, for each circumstance the parties might find themselves in, the assignment of bargaining power is specified in the contract, rather than being given exogenously. Aghion-Dewatripont-Rey do not address what happens when parties abandon their current contract, which is what we mean by renegotiation.

where $h(a,\theta)$ is the equilibrium renegotiated outcome, starting from the mechanism-prescribed alternative a in state θ. We shall always make the following three assumptions about $h(\cdot,\cdot)$.

Assumption A1 (Renegotiation is predictable): $h(\cdot,\cdot)$ is a function that is common knowledge to the individuals;

Assumption A2 (Renegotiation is efficient): $h(a,\theta)$ is Pareto optimal for all $a \in A$ and $\theta \in \Theta$ (that is, there does not exist $b \in A$ such that $bR_i(\theta)h(a,\theta)$ for all i, with strict preference for some i);

Assumption A3 (Renegotiation is individually rational): For all $a \in A$ and $\theta \in \Theta$, and all i, $h(a,\theta)R_i(\theta)a$.

Of these assumptions, A3 is perhaps the least controversial: no individual need be forced into a renegotiation process that is going to make him worse off, since he could always insist on abiding by the alternative specified by the mechanism.

Assumption A2 is strong but reasonable in our framework of ex post complete information. If individuals anticipated that the renegotiation process were going to result in an inefficient outcome, then one of them could presumably propose a Pareto improvement that the others would accept. In any case, if renegotiation resulted in inefficient outcomes, that would

only make the implementation problem easier (since it would worsen the penalty of out-of-equilibrium outcomes). Thus, because we are interested in the extent to which renegotiation constrains implementation, it is natural to tie our hands as much as possible.

As for A1, we note that obviously renegotiation results in some particular outcome, and so it is the hypothesis that agents can predict what this outcome will be and that this prediction is common knowledge (rather than the insistence that h be a function) that is the restrictive aspect of the assumption. In effect we are assuming that the state θ -- in addition to resolving uncertainty about preferences -- also resolves uncertainty about the outcome of any future bargaining amongst the agents.

This assumption does not necessarily mean that individuals can forecast ex ante how renegotiation will proceed. Before the state is realized, there might be considerable uncertainty about relative bargaining power, etc. Indeed, although we have so far been emphasizing the differences in preferences that different states could entail, our model allows for the possibility that two states θ and ϕ might be identical in preferences and differ only in terms of how renegotiation would proceed.

Suppose, contrary to A1, that there is ex post uncertainty about renegotiation, i.e., that we can write the equilibrium renegotiated outcome as a random variable $\tilde{h}(a,\theta)$. Then, as long as agents have common beliefs about the distribution of $\tilde{h}(a,\theta)$, this case is not conceptually different from that of certainty. Indeed, as with inefficient renegotiation, uncertainty about the realization of $\tilde{h}(a,\theta)$ may actually facilitate implementation rather than impede it. This is because even though (from A2)

each realization of $\tilde{h}(a,\theta)$ is Pareto optimal (i.e., it lies on the Pareto frontier of the utility possibility set), the expected utilities from $\tilde{h}(a,\theta)$ (which correspond to a convex combination of the utilities from each realization) may lie in the interior of the utility possibility set. Thus the "point expectation" assumption embodied in A1 actually only makes implementation harder.

(We should remark that because randomization over Pareto optima may itself not be Pareto optimal, lotteries can be exploited by a mechanism-designer. That is, in constructing an implementing mechanism, one can deliberately introduce randomization as an effective way of punishing agents for deviations. This important idea will be amplified in Section 3.)

Let us give two examples of renegotiation functions $h(\cdot,\cdot)$ that have been used frequently in the literature. For this purpose, it is useful to represent each agent i's preferences in state θ by the von Neumann-Morgenstern utility function $u_i(\cdot,\theta)$.

Monopolistic Bargaining Power: In this case, some agent j has all the bargaining power in renegotiation. That is, he can credibly make take-it-or-leave-it offers to the other agents, and so he derives all the surplus from bargaining. This means that

$$h(a,\theta) = \underset{b}{\text{argmax}}\ u_j(b,\theta) \text{ subject to } u_i(b,\theta) \geq u_i(a,\theta) \text{ for all } i \neq j.$$

Nash Bargaining: In this case, the Nash bargaining solution applies to agents' renegotiations:

$$h(a,\theta) = \underset{b}{\operatorname{argmax}} \prod_{i=1}^{n} (u_i(b,\theta) - u_i(a,\theta)).$$

Notice that both these examples satisfy assumptions A1-A3.

We can now define more formally what it means for an SCR to be implementable when renegotiation is possible. To do so, it will be convenient henceforth to restrict attention to SCRs f that are essentially single-valued, in the sense that all individuals are indifferent between any two f-optimal alternatives.

Assumption A0 ("Single-valuedness" of f): For all $\theta \in \Theta$ and all $a \in f(\theta)$,

$$f(\theta) = \{b \in A \mid aI(R_i(\theta))b \text{ for all } i\}.$$

Given an SCR f (satisfying A0), a renegotiation function h, and an equilibrium concept, we will say that f is implementable in that equilibrium concept for renegotiation function h if there exists a mechanism g such that

(**) $\quad \varnothing \neq EQ_{hog}(\theta) \subseteq f(\theta)$ for all $\theta \in \Theta$.[9]

Here "hog" denotes the composition of h and g. That is, if g results in an outcome a in state θ, the outcome under hog is $h(a,\theta)$,

Our definition of implementation with renegotiation suggests that, qualitatively, there should be little difference between implementation theory when renegotiation is possible and the standard theory where renegotiation is ruled out. Indeed, it would seem that obtaining a result for renegotiation would be a matter of "translating" the corresponding result from the standard framework. That is, take the standard result and "apply h" to it.

Formally, this translation principle is correct. If, for instance, Nash equilibrium is the equilibrium concept, then we know from standard theory

---

[9]Note that (**) is a somewhat weaker notion of implementation than (*) in the Introduction, where we demanded <u>equality</u> between the set of equilibrium outcomes and $f(\theta)$. Now we require only that the set of equilibrium outcomes (for the composite game hog) be nonempty and a subset of $f(\theta)$. The reason for this relaxation is that we could have a situation in which a and b are optimal in state θ (i.e., $a,b \in f(\theta)$) and all agents are indifferent between a and b, but the renegotiation function h only renegotiates to a, never b. (In particular, $h(b,\theta) = a$.) So we could never actually have b as an equilibrium outcome.) As long as we impose A0, however, this difficulty makes no difference to agents' utilities.

(see Maskin (1999)) that monotonicity of an SCR is a necessary and almost sufficient condition for its implementability.[10] The SCR is <u>monotonic</u> provided that, for all $\theta \in \Theta$ there exists $a \in A$ such that $a \in f(\theta)$ and such that, for all $\phi \in \Theta$, if, for all i and all $b \in A$, the implication $[aR_i(\theta)b \longrightarrow aR_i(\phi)b]$ holds, then $a \in f(\phi)$. That is, for each state, there exists an f-optimal alternative a such that if we now change preferences so that, in each agent's preference ordering, a does not fall below any alternative it was not ranked below before, then a remains f-optimal.[11]

The translation principle suggests that the following "translated" condition -- <u>renegotiation-monotonicity</u> -- should be the key to implementability when renegotiation is possible. An SCR f is renegotiation-monotonic for renegotiation function h provided that, for all $\theta \in \Theta$, there exists $a \in A$ such that $h(a,\theta) \in f(\theta)$ and such that, for all $\phi \in \Theta$, if, for all i and all $b \in A$, the implication $[h(a,\theta)R_i(\theta)h(b,\theta) \longrightarrow h(a,\phi)R_i(\phi)h(b,\phi)]$ holds, then $h(a,\phi) \in f(\phi)$. And indeed Theorem 5 below establishes formally that renegotiation-monotonicity <u>is</u> the key.

---

[10] More precisely, monotonicity together with a weak condition ("no veto power") is sufficient for implementability in the case $n \geq 3$ (no veto power says that if all agents but one top-rank an alternative a, then a must be f-optimal).

[11] This version of monotonicity is actually slightly weaker than the usual version because we are using a slightly weaker-than-usual concept of implementation. (Recall (**) and the discussion in footnote 9.)

But translations, such as Theorem 5, of standard results from the literature are too abstract to give a clear indication of how serious a constraint renegotiation is for actual models studied in the renegotiation literature. In principle, there are two sorts of ways in which renegotiation can make implementation harder. The first, which we already mentioned in the Introduction, is that renegotiation may make the threat of harsh consequences for deviating less credible (because those consequences may be renegotiated away). We saw an illustration of this in the example of Table 1: agent 2 got stuck with the inferior alternative b if she failed to choose a in state θ. But if b then got renegotiated to c, it no longer served as an effective deterrent.

The other problem that renegotiation poses is that it may interfere with "preference reversal." If all agents' preferences are the same in states θ and ϕ, then (in the absence of renegotiation) it is clearly impossible to design a game form in which the equilibrium outcomes are different in the two states; strategically, the two states are equivalent. Put another way, if it is desirable that a be implemented in state θ and b in state ϕ, then there had better be at least one agent for whom preferences are different (not necessarily between a and b) in the two states, i.e., for whom there is preference reversal. To return to the example of Table 1, notice that agent 2's preferences over a and b reverse between the states θ and ϕ: he prefers a to b in state θ, and b to a in ϕ. And this difference is the key to why the simple mechanism of his choosing between a to b works (in the absence of renegotiation). Notice, however, that if b is renegotiated to c in state θ, then Agent 2 prefers $h(b,\xi)$ to $h(a,\xi)$ for both $\xi = \theta$ and $\xi = \phi$, i.e., there is no longer preference reversal over a and b; renegotiation has the effect

of "destroying" it.

The example of Table 1 is too crude to distinguish between these two problems that renegotiation can create. In Section 4, we introduce a richer example consisting of a buyer-seller contractual relationship (similar to many studied in the incomplete contracts literature). We revisit this example after each of Theorems 1-4, and we will come to the conclusion that it is the <u>former</u> problem -- ineffectiveness of punishment -- rather than the destruction of preference reversal that is typically the more serious problem.

3. <u>Lotteries</u>

As remarked in Section 2, it may be desirable in some circumstances to deliberately introduce randomizations over alternatives. Accordingly, let ΔA be the set of all random variables (lotteries) over A. We shall denote a typical element of ΔA by $\tilde{a}$. And, henceforth, we shall interpret agent i's preference ordering $R_i(\theta)$ as an ordering over lotteries; we assume this ordering can be represented by a von Neumann-Morgenstern utility function. The (nonstochastic) choice rule f is Pareto optimal with respect to ΔA: for all $\theta \in \Theta$ and all $a \in f(\theta)$, there does not exist $\tilde{b} \in \Delta A$ such that $\tilde{b} R_i(\theta) a$ for all i, with strict preference for some i.

The reader may wonder how lotteries could make a difference. Wouldn't they simply be renegotiated away? What we have in mind is that if, say, two agents play strategies (s_1, s_2) of a mechanism g for which $g(s_1, s_2) = \tilde{a}$, the randomization is performed mechanically and <u>instantaneously</u>. In this way,

there is no time to renegotiate. Of course, there can be renegotiation after the randomization occurs. Hence, following the play (s_1,s_2), the agents face a stochastic final outcome $h(\tilde{a},\theta)$, where $h(\tilde{a},\theta)$ denotes the random variable that corresponds to the renegotiations from all possible realizations of $\tilde{a}$. The crucial point is that although $h(\tilde{a},\theta)$ is Pareto optimal for each realization of $\tilde{a}$, the lottery $h(\tilde{a},\theta)$ need not be Pareto optimal if the parties are risk averse.

What about renegotiation before (s_1,s_2) is played? If (s_1,s_2) is an out-of-equilibrium configuration, then agents do not anticipate that (s_1,s_2) will be played, and so have no need to renegotiate $\tilde{a}$. Indeed, one of them would necessarily be strictly worse off if the renegotiation occurred and thereby destroyed equilibrium (since someone is always made worse off in moving from one Pareto optimum to another), and so he will resist renegotiating. On the other hand, if (s_1,s_2) constitutes an equilibrium, then agents would renegotiate $\tilde{a}$ before the strategies are played, unless the lottery $h(\tilde{a},\theta)$ itself were Pareto optimal. Thus we will need to impose the constraint that if $\tilde{a}$ occurs in equilibrium (or in a continuation equilibrium, if the mechanism is a multistage game), then the lottery $h(\tilde{a},\theta)$ must be Pareto-optimal in state θ.

Whenever we write $h(\tilde{a},\theta) \in f(\theta)$, we mean that the lottery $h(\tilde{a},\theta)$ is f-optimal in state θ. This could occur either if agents are indifferent over all realizations of $h(\tilde{a},\theta)$ (and all of these are f-optimal), or if the realizations lie along a linear portion of the Pareto frontier, in which case the realizations themselves may not be f-optimal, but their expectation is.

4. The Case of Two Agents

Let us begin our formal analysis with the case of two agent (n = 2). This is the most pertinent case for the large literature on bilateral contracts.

We can readily develop a set of conditions that are necessary for implementation with renegotiation regardless of the refinement of Nash equilibrium that is adopted as the solution concept (that is, whether the concept is Nash equilibrium or any of its refinements, the necessary conditions are the same):

Theorem 1: Assume that n = 2. Then the SCR f can be implemented in Nash equilibrium (or in any refinement of Nash equilibrium, e.g., subgame-perfect equilibrium) for renegotiation function h only if

(I) there exists a random function $\tilde{a}(\cdot,\cdot)$: $\Theta \times \Theta \rightarrow \Delta A$ such that, for all $\theta \in \Theta$, $h(\tilde{a}(\theta,\theta),\theta) \in f(\theta)$; and

(II) for all $\theta,\phi \in \Theta$,

$$h(\tilde{a}(\theta,\theta),\theta)R_1(\theta)h(\tilde{a}(\phi,\theta),\theta) \tag{1}$$

and

$$h(\tilde{a}(\theta,\theta),\theta)R_2(\theta)h(\tilde{a}(\theta,\phi),\theta). \tag{2}$$

Proof: Suppose that there exists a game form, expressed in normal form as $g: S_1 \times S_2 \rightarrow \Delta A$, that implements f. Then for states θ and ϕ there exist equilibria $(s_1(\theta), s_2(\theta))$ (in state θ) and $(s_1(\phi), s_2(\phi))$ (in state ϕ) such that

$$h(g(s_1(\xi), s_2(\xi)), \xi) \in f(\xi), \text{ for } \xi = \theta, \phi. \tag{3}$$

For all ξ and ξ', define $\tilde{a}(\xi, \xi') = g(s_1(\xi), s_2(\xi'))$. Then hypothesis I follows from (3). (1) follows from the fact that, in state θ, agent 1 must weakly prefer $s_1(\theta)$ to $s_1(\phi)$. (2) follows from the fact that in state θ agent 2 must weakly prefer playing $s_2(\theta)$ to $s_2(\phi)$.

Q.E.D.

Although very simple, Theorem 1 is remarkably useful in narrowing down the set of implementable SCRs in given models.

Consider first the example of Table 1 when $h(b, \theta) = c$. We claim that there is no way of implementing f such that $f(\theta) = a$ and $f(\phi) = b$ for implementation function h. Suppose, to the contrary, that $\tilde{a}(\cdot, \cdot)$ satisfying hypotheses I and II existed. From hypothesis I, $h(\tilde{a}(\theta, \theta), \theta) = a$ and $h(\tilde{a}(\phi, \phi), \phi) = b$. Since $h(b, \theta) = c$, it follows from (2) and A3 that $\tilde{a}(\theta, \phi)$ must equal a. But then, by A3, $h(\tilde{a}(\theta, \phi), \phi) P(R_1(\phi)) h(\tilde{a}(\phi, \phi), \phi)$, which, reversing the roles of θ and ϕ, contradicts (1).

Consider next the following buyer-seller example, which is typical of many models in the incomplete contracts literature. A buyer and seller, who are both risk-neutral, wish to trade an item which the seller currently

possesses. The good is worth nothing to the seller, but she must decide whether to undertake a prior investment, at a private cost (in money) to her of c, that enhances the value of the good to the buyer. If the investment is undertaken (state θ), the good is worth $\bar{v}$ to the buyer; otherwise it is worth $\underline{v}$ (state ϕ), where $\bar{v} > \underline{v} > 0$. Assume that the gain, $\bar{v} - \underline{v}$, exceeds the cost, c, so that it is efficient for the seller to invest. The interesting case is where the gain is less than twice the cost: i.e., where c lies between $\frac{1}{2}(\bar{v} - \underline{v})$ and $\bar{v} - \underline{v}$. The question then is: Can a mechanism be designed (and specified in a contract that both parties sign beforehand) which induces the seller to invest, if both the investment and the state are unverifiable?

Notice that, in either state (i.e. whether or not the seller has sunk the cost of investment), trade will always take place because the good is worth more to the buyer than to the seller ($\bar{v}, \underline{v} > 0$), and the parties can renegotiate whatever mechanism has been contractually agreed.

To see the effect of renegotiation, consider the following contract: a price, p, is contracually specified at which the seller must supply the good, but the buyer the right to refuse delivery (and in which case, he pays nothing). The idea is that by fixing a high price, say $p = \bar{v} - \varepsilon$ where $\varepsilon > 0$ is small, the buyer won't accept the good unless the investment has been made, and that the seller will get a positive return p - c from making the investment. Unfortunately, renegotiation undoes such a contract. Even when the seller has invested (state θ), the buyer can strategically refuse delivery of the good, and then, outwith the contract, renegotiate the trading price down from p. If we assume Nash bargaining (marginal surplus is divided equally), then the renegotiated price will be $\bar{v}/2$. Moreover, if the seller hasn't invested (state ϕ), then, after the buyer refuses delivery, the

parties renegotiate and trade at a price $\underline{v}/2$. Since the seller's gain from investment, $\frac{1}{2}(\bar{v} - \underline{v})$, is less than her cost, c, she will not invest.

Here, renegotiation actually makes the contract redundant: the parties might as well have written no contract. (With no contract, the trading price would still be $\bar{v}/2$ in state θ and $\underline{v}/2$ in state ϕ.) In general, however, contracts <u>will</u> make a difference, so the question remains: Is there a contract that induces the seller to invest?

For any contract let $p(\xi)$ denote the (expected) total amount the buyer pays the seller in state $\xi = \theta,\phi$. Clearly, the seller will choose to invest only if the price difference, $p(\theta) - p(\phi)$, exceeds her cost, c.

Let the buyer and seller correspond to agents 1 and 2 in Theorem 1, respectively. We know from (1) that there must exist some (stochastic) alternative $\tilde{a}(\phi,\theta)$ such that $\bar{v} - p(\theta) \geq \bar{v} - Eh(\tilde{a}(\phi,\theta),\theta)$ -- where, with a slight abuse of notation, $Eh(\tilde{a}(\phi,\theta),\theta)$ denotes the expected total amount the buyer pays the seller following the renegotiation of $\tilde{a}(\phi,\theta)$ in state θ. (Notice that we can take expectations here, given that the buyer is risk neutral.) Also, reversing the roles of θ and ϕ, we know from (2) that $p(\phi) \geq Eh(\tilde{a}(\phi,\theta),\phi)$. (Again, we can take expectations, given that the seller is risk neutral.) Hence the price difference $p(\theta) - p(\phi)$ cannot exceed $Eh(\tilde{a}(\phi,\theta),\theta) - Eh(\tilde{a}(\phi,\theta),\phi)$.

Without loss of generality, let the alternative $\tilde{a}(\phi,\theta)$ be represented by the following lottery: with some specified probability λ, the parties trade and the buyer pays the seller an amount p_1; and with probability $1 - \lambda$, the parties do not trade (at least not prior to renegotiation) and the buyer pays

the seller p_0. Whenever "no trade" is specified there will be renegotiation, and, given Nash bargaining, the buyer will pay the seller (over and above p_0) either an additional amount $\frac{1}{2}\bar{v}$ (in state θ), or an additional amount $\frac{1}{2}\underline{v}$ (in state ϕ). Hence $Eh(\tilde{a}(\phi,\theta),\theta) = \lambda p_1 + (1 - \lambda)(p_0 + \frac{1}{2}\bar{v})$ and $Eh(\tilde{a}(\phi,\theta),\phi) = \lambda p_1 + (1 - \lambda)(p_0 + \frac{1}{2}\underline{v})$. This means that the price difference $p(\theta) - p(\phi)$ cannot exceed $\frac{1}{2}(1 - \lambda)(\bar{v} - \underline{v})$, which in turn is less than $\frac{1}{2}(\bar{v} - \underline{v})$.

Theorem 1 thus enables us reach the negative conclusion that whenever the gain from investment is less than twice the cost (i.e., whenever c lies between $\frac{1}{2}(\bar{v} - \underline{v})$ and $\bar{v} - \underline{v}$)), there is no contract that will give the seller an incentive to invest, even though investment is efficient. The intuition is general: with renegotiation, no contract can recoup the seller more than fifty cents of every dollar of benefit that she bestows on the buyer, and this externality dilutes her incentive to incur the private cost of investment.

One salient feature of this example is that, in each state, the Pareto frontier of the utility possibility set is linear. That is, in state θ, the equation of the Pareto frontier (net of the seller's cost) is $u_1 + u_2 = \bar{v} - c$ where u_i is agent i's payoff; and in state ϕ the equation is $u_1 + u_2 = \underline{v}$. This linearity is a feature that is shared by virtually all models in the incomplete contracts literature, and follows from the facts that agents are risk-neutral and utilities are quasi-linear (linear in money and additively separable).

A striking implication of linearity is that the hypotheses of Theorem 1 are not only necessary but also <u>sufficient</u> for implementation. One way to construe condition (1) is that in state θ agent 1 can be punished with

outcome $\tilde{a}(\phi,\theta)$ for deviating from equilibrium. Likewise, condition (2) says that agent 2 can be punished with outcome $\tilde{a}(\theta,\phi)$. But as we noted in section 2, in addition to punishment, one needs preference reversal to implement an SCR. It turns out that, with linearity, the fact that either agent can be punished implies that preference reversal exists.

To establish this, we need a general definition of linearity. For all states θ, we shall call the Pareto frontier in state θ linear if, for all $a, b \in A$ such that a and b are Pareto optimal in state θ, any lottery $\lambda a + (1 - \lambda)b$ between a and b (where λ is the probability of a and $1 - \lambda$ is the probability of b) is also Pareto optimal.

Theorem 2: Assume that $n = 2$ and the Pareto frontier is linear in all states $\theta \in \Theta$. Then the SCR f can be implemented in Nash equilibrium (or any refinement) for renegotiation function h if there exists a function $\tilde{a}(\cdot,\cdot)$: $\Theta \times \Theta \longrightarrow \Delta A$ satisfying hypotheses I and II of Theorem 1.

Proof: Consider the following mechanism g. Each agent i announces a state $\theta^i \in \Theta$, and the outcome of g, given the agents' announcements, is defined to be the (possibly random) alternative $\tilde{a}(\theta^1,\theta^2)$, where $\tilde{a}(\cdot,\cdot)$ is the function specified in the hypotheses of the theorem.

Suppose that the true profile is θ. From hypothesis II, for each agent i it is an equilibrium of the composite game $h \circ g$ to set $\theta^i = \theta$. Moreover, from hypothesis I the equilibrium outcome is f-optimal. Finally, the

linearity of the Pareto frontier implies that the composite game $h \circ g$ is zero-sum. Hence, from assumption A0, any other equilibrium in state θ is also f-optimal. We conclude that g implements f with renegotiation function h.

Q.E.D.

Although linearity of the Pareto frontier is often assumed in the literature, it is nevertheless a quite restrictive assumption. In particular, we will see below that it has much to do with the negative conclusion we reached in our earlier buyer-seller example.

We first establish a counterpart to Theorem 2 in the case where the Pareto frontier is not necessarily linear:

Theorem 3: Assume that n = 2. The SCR f can be implemented in subgame-perfect equilibrium with renegotiation function h if:

(III) there exist random function $\tilde{a}(\cdot)$: $\Theta \rightarrow \Delta A$ such that, for all $\theta \in \Theta$, $h(\tilde{a}(\theta),\theta) \in f(\theta)$;

(IV) for all $\theta,\phi \in \Theta$ such that $h(\tilde{a}(\theta),\phi) \notin f(\phi)$ there exists an agent $k(\theta,\phi)$ and a pair of alternatives $\tilde{b}(\theta,\phi)$ and $\tilde{c}(\theta,\phi)$ in ΔA such that

$$h(\tilde{b}(\theta,\phi),\theta)R_{k(\theta,\phi)}(\theta)h(\tilde{c}(\theta,\phi),\theta) \tag{4}$$

and

$$h(\tilde{c}(\theta,\phi),\phi)P(R_{k(\theta,\phi)}(\phi))h(\tilde{b}(\theta,\phi),\phi); \tag{5}$$

(V) if $X \subseteq \Delta A$ is the union of all $\tilde{a}(\xi)$ for $\xi \in \Theta$, together with all $\tilde{b}(\xi,\xi')$ and $\tilde{c}(\xi,\xi')$ for $\xi,\xi' \in \Theta$ (when these are defined), then no alternative in X is ever maximal for any agent i in any state $\theta \in \Theta$, even after renegotiation h (that is, there exists some $d^i(\theta) \in \Delta A$ such that $d^i(\theta)P(R_i(\theta))h(x,\theta)$ for all $x \in X$); and

(VI) there exists some alternative $\tilde{e} \in \Delta A$ such that, for any agent i in any state $\theta \in \Theta$, every alternative in X is strictly preferred to $\tilde{e}$ after renegotiation h (that is, $h(x,\theta)P(R_i(\theta))h(\tilde{e},\theta)$ for all $x \in X$).

Proof: We start with some definitions. First, given that the lottery $h(\tilde{b}(\theta,\phi),\theta)$ may not be Pareto optimal in state θ, define $\hat{h}(\tilde{b}(\theta,\phi),\theta)$ to be the (Pareto optimal) outcome that is reached if renegotiation occurs before the resolution of the randomness in $\tilde{b}(\theta,\phi)$. Because $h(\tilde{a}(\theta),\theta)$ is Pareto

optimal (from hypothesis III), there exists, for all ϕ such that $h(\tilde{a}(\theta),\phi) \notin f(\phi)$, an agent $j(\theta,\phi)$ such that

$$h(\tilde{a}(\theta),\theta)R_{j(\theta,\phi)}(\theta)\hat{h}(\tilde{b}(\theta,\phi),\theta). \tag{6}$$

Next, from hypothesis VI, we can define an alternative $\tilde{q} \in \Delta A$ such that, for all $\theta \in \Theta$ and all $x \in X$,

$$h(x,\theta)P(R_i(\theta))h(\tilde{q},\theta)P(R_i(\theta))h(\tilde{e},\theta) \quad \text{for } i=1,2. \tag{7}$$

(For example, we can take $\tilde{q}$ to be a randomization between $\tilde{e}$ and some alternative in X.)

Finally, for $i = 1,2$ and all $\theta \in \Theta$, define $\tilde{r}^i(\theta)$ to be agent i's favorite alternative in ΔA when θ is the state, subject to the constraint that the other agent, $-i$, would not want to veto it in favor of $\tilde{q}$. There are two cases to consider. Either the constraint is binding, in which case agent $-i$ is indifferent between $\tilde{r}^i(\theta)$ and $h(\tilde{q},\theta)$, $\tilde{r}^i(\theta)$ is Pareto optimal, and from (7) it follows that, for all $x \in X$,

$$\tilde{r}^i(\theta)P(R_i(\theta))h(x,\theta).^{12} \tag{8}$$

---

[12] If there is no nonstochastic Pareto optimal alternative $r^i(\theta)$ such that $r^i(\theta)I(R_{-i}(\theta))h(\tilde{q},\theta)$, then agent i can choose a lottery $\tilde{r}^i(\theta)$ of nonstochastic Pareto optimal alternatives to push agent $-i$ down to the point of indifference. Moreover, this lottery can be chosen so that $\tilde{r}^i(\theta)$ is also Pareto optimal.

Or the constraint is not binding, in which case without loss of generality $\tilde{r}^i(\theta)$ can be chosen to be Pareto optimal, and (8) follows directly from hypothesis V.

Given these definitions, we can now proceed to construct a stage mechanism that implements f with renegotiation function h. In the first stage, agent 2 announces two states, θ^2 and ϕ^2, and a non-negative integer m^2. Simultaneously, agent 1 announces three mappings, $\Lambda^1: \Theta \rightarrow \Theta$, $\Gamma^1: \Theta \rightarrow \Theta$, and $M^1: \Theta \rightarrow \{0,1,2,\ldots\}$. If $\Lambda^1(\theta^2) \neq \theta^2$, then the outcome of the mechanism is $\tilde{e}$. If, instead, $\Lambda^1(\theta^2) = \theta^2$, then what happens depends on Γ^1, M^1, ϕ^2, and m^2. Let $\phi^1 = \Gamma^1(\theta^2)$ and $m^1 = M^1(\theta^2)$. If $m^1 = m^2 = 0$, the outcome of the mechanism is $\tilde{a}(\theta^2)$. If, for some j, $m^j > \min\{m^1,m^2\} = 0$, then the outcome of the mechanism is still $\tilde{a}(\theta^2)$ provided that $h(\tilde{a}(\theta^2),\phi^j) \in f(\phi^j)$ or $j \neq j(\theta^2,\phi^j)$ (where $j(\theta^2,\phi^j)$ is defined in (6) above); otherwise the mechanism moves to stage $2(\theta^2,\phi^j)$ (see below). Finally, if $\min\{m^1,m^2\} > 0$, the agent who has announced the highest integer from m^1,m^2 (with ties broken by a coin flip) gets to choose any alternative in ΔA.

In stage $2(\theta^2,\phi^j)$ (which, as constructed above, is reached only if $\Lambda^1(\theta^2) = \theta^2$, $h(\tilde{a}(\theta^2),\phi^j) \notin f(\phi^j)$, and the agent j announcing ϕ^j is such that $j = j(\theta^2,\phi^j)$ and $m^j > \min\{m^1,m^2\} = 0$), agent 2 announces a non-negative integer n^2 and agent 1 announces a mapping $N^1: \{0,+\} \rightarrow \{0,1,\cdots\}$. If $N^1(0) = n^2 = 0$ then agent $j(\theta^2,\phi^j)$ gets to choose any alternative in ΔA, subject to the other agent's veto; if the veto is exercised, the outcome is $\tilde{q}$. If either ($n^2 = 0$ and $N^1(0) > 0$) or ($n^2 > 0$ and $N^1(+) = 0$), then the outcome of the mechanism is $\tilde{e}$. Next let $n^1 = N^1(+)$ and suppose that $\min\{n^1,n^2\} = 1$. If $n^{k(\theta^2,\phi^j)} = 1$ (where $k(\theta^2,\phi^j)$ is defined in hypothesis IV), then the outcome

of the mechanism is $\tilde{b}(\theta^2,\phi^j)$, whereas if $n^{k(\theta^2,\phi^j)} > 1$, the outcome is $\tilde{c}(\theta^2,\phi^j)$. Finally, if $\min\{n^1,n^2\} > 1$, the agent who has announced the highest integer from n^1,n^2 (with ties broken by a coin flip) gets to choose any alternative in ΔA.

Let us verify that this mechanism implements f with renegotiation function h. Suppose throughout that θ is the true state.

We will demonstrate that the following is a subgame-perfect equilibrium, whose outcome, following renegotiation, is $h(\tilde{a}(\theta),\theta)$ -- which, by hypothesis III, is contained in $f(\theta)$. At the first stage, agent 2 sets $(\theta^2,\phi^2,m^2) = (\theta,\theta,0)$; agent 1 sets $\Lambda^1(\cdot) \equiv \theta$, $\Gamma^1(\cdot) \equiv \theta$, and $M^1(\cdot) \equiv 0$. In a subgame where $\Lambda^1(\theta^2) = \theta^2$ and $\min\{M^1(\theta^2),m^2\} > 0$, the agent setting the highest integer from $M^1(\theta^2),m^2$ chooses his favorite alternative in ΔA. If play reaches stage $2(\theta,\phi^j)$, agent 2 sets $n^2 = 1$ and agent 1 sets $N^1(\cdot) \equiv 1$. If play reaches stage $2(\theta^2,\phi^j)$ where $\theta^2 \neq \theta$, agent 2 sets $n^2 = 0$ and agent 1 sets $N^1(\cdot) \equiv 0$. In a subgame where $N^1(0) = n^2 = 0$, agent $j(\theta^2,\phi^j)$ chooses $\tilde{r}^{j(\theta^2,\phi^j)}$, and the other agent exercises his veto if only if he strictly prefers $h(\tilde{q},\theta)$ to agent $j(\theta^2,\phi^j)$'s choice. Finally, in a subgame where $\min\{N^1(+),n^2\} > 1$, the agent setting the highest integer from $N^1(+),n^2$ chooses his favorite alternative in ΔA.

To see that this is a subgame-perfect equilibrium, let us work backward from the end of the game. The specified behaviors in the subgames after $N^1(0) = n^2 = 0$ and after $\min\{N^1(+),n^2\} > 1$ are clearly optimal. Now if agents set $N^1(\cdot) \equiv n^2 = 1$ at stage $2(\theta,\phi^j)$, the outcome is $\tilde{b}(\theta,\phi^{j(\theta,\phi^j)})$. If instead agent $k(\theta,\phi^j)$ deviates by choosing an integer greater than 1 or else zero, he induces $\tilde{c}(\theta,\phi^{j(\theta,\phi^j)})$ or $\tilde{e}$, respectively. But from (4) and

hypothesis VI neither deviation is profitable. Similarly, the other agent can induce $\tilde{e}$ by deviating to zero, but from hypothesis VI does not gain from doing so. Hence the specified second-stage behavior at stage $2(\theta,\phi^j)$ constitutes a continuation equilibrium. What about the specified behavior at stage $2(\theta^2,\phi^j)$ where $\theta^2 \neq \theta$? If neither agent deviates from $N^1(\cdot) \equiv n^2 = 0$, the outcome in the ensuing subgame is $\tilde{r}^{j(\theta^2,\phi^j)}(\theta)$. But if either agent deviates unilaterally, the only possible alternative outcome is $\tilde{e}$, which from (7) and (8) is strictly inferior for both agents. Thus again the specified behavior forms an equilibrium.

Moving back to the first stage, we note that if agents adhere to the specified strategies then the outcome is $\tilde{a}(\theta)$. Now if one of the agents deviates unilaterally, the only possible continuation equilibrium outcome other than $\tilde{a}(\theta)$ is either $\tilde{e}$ (if $\Lambda^1(\theta^2) \neq \theta^2$), which from hypothesis VI is worse than $\tilde{a}(\theta)$, or $\tilde{b}(\theta,\phi^j)$ (this outcome arises if stage $2(\theta,\phi^j)$ is reached and agent $j(\theta,\phi^j)$ is the deviant), in which case (6) implies that agent $j(\theta,\phi^j)$ again does not gain from deviating, even if $\tilde{b}(\theta,\phi^j)$ is renegotiated before its randomness is resolved.

This completes the demonstration that there exists an equilibrium with outcome $\tilde{a}(\theta)$ such that $h(\tilde{a}(\theta),\theta) \in f(\theta)$. It remains to show that all other equilibrium outcomes are also in $f(\theta)$ (after renegotiation).

We start with the claim that at stage $2(\theta^2,\phi^j)$, the only possible continuation equilibrium outcomes are $\tilde{b}(\theta^2,\phi^j)$ and $\tilde{r}^{j(\theta^2,\phi^j)}(\theta)$. Moreover, if $\phi^j = \theta$, then $\tilde{r}^{j(\theta^2,\phi^j)}(\theta)$ is the unique continuation equilibrium outcome.

To prove this claim, first note that if $n^2 = 0$ with positive

probability, then agent 1's optimal response is to set $N^1(0) = 0$, and the outcome is $\tilde{r}^{j(\theta^2,\phi^j)}(\theta)$: otherwise the outcome will be $\tilde{e}$, which from (7) and (8) is strictly inferior. Next, note that if $n^2 \geq 1$ with positive probability, then agent 1 will never set $N^1(+) = 0$, since this is strictly dominated by $N^1(+) = 1$: from hypothesis VI, $\tilde{e}$ is strictly dominated by both $\tilde{b}(\theta^2,\phi^j)$ and $\tilde{c}(\theta^2,\phi^j)$.

Now let k be agent $k(\theta^2,\phi^j)$, and -k be the other agent. If $n^2 \geq 1$ with positive probability, then, conditional on $n^2 \geq 1$, it is clear that the only candidate continuation equilibrium of the "integer game" $\{n^1,n^2\}$ has $n^k = 1$ and $n^{-k} \geq 1$, with outcome $\tilde{b}(\theta^2,\phi^j)$. (The reason is that if $n^2 \geq 1$ with positive probability, then $N^1(+) \geq 1$. But if there is a positive probability of agent k announcing $n^k \geq 2$, then, for agent -k, announcing $n^{-k} = 1$ is, by hypothesis V, strictly dominated by announcing a suitably large n^{-k}, so as to implement his favorite alternative (with arbitrarily small risk of being outbid by agent k) whenever $n^k \geq 2$, without changing the outcome when $n^k = 1$. But then, given that agent -k never announces $n^{-k} = 1$, agent k should try to win the integer game himself; and clearly there is no equilibrium of this kind. Hence $n^k = 1$ and $n^{-k} \geq 1$, conditional on $n^2 \geq 1$.) Finally, notice that if $\phi^j = \theta$, then conditional on $n^2 \geq 1$, $n^k = 1$ cannot be part of a continuation equilibrium either, because, by announcing a suitably large n^k, agent k can induce $\tilde{c}(\theta^2,\theta)$ or his favorite alternative with arbitrarily high probability, which, from (5) and hypothesis V, strictly dominate $\tilde{b}(\theta^2,\theta)$. Hence the claim is established.

Let us move back to the first stage. If agent 2 announces some θ^2 with positive probability, then $\Lambda^1(\theta^2) = \theta^2$ with probability 1, since any other value of $\Lambda^1(\theta^2)$ (and any $M^1(\theta^2)$) is strictly dominated by $\Lambda^1(\theta^2) = \theta^2$, $M^1(\theta^2)$

= 0: by (7) and (8), $\tilde{e}$ is strictly dominated both by $\tilde{a}(\theta^2)$ and by any possible equilibrium outcome of the second stage ($\tilde{b}(\theta^2,\phi^j)$ or $\tilde{r}^{j(\theta^2,\phi^j)}(\theta)$).

Conditional on $\Lambda^1(\theta^2) = \theta^2$, it is clear that the only candidate equilibrium of the "integer game" $\{m^1,m^2\}$ is $m^1 = m^2 = 0$ with outcome $\tilde{a}(\theta^2)$. (The reason is that if, for some j, agent -j announces $m^{-j} \geq 1$ with positive probability, then, for agent j, announcing $m^j = 0$ (and any ϕ^j) is strictly dominated by announcing a suitably large m^j together with $\phi^j = \theta^2$ (implying $h(\tilde{a}(\theta^2),\phi^j) \in f(\phi^j)$), so as to implement his favorite alternative (with arbitrarily small risk of being outbid by agent -j) whenever $m^{-j} \geq 1$, without changing the outcome when $m^{-j} = 0$. But then, given that agent j never announces $m^j = 0$, agent -j should try to win the integer game himself; and clearly there is no equilibrium of this kind.) However, if $h(\tilde{a}(\theta^2),\theta) \notin f(\theta)$, then we have shown that the unique equilibrium outcome of stage $2(\theta^2,\theta)$ is $\tilde{r}^{j(\theta^2,\theta)}(\theta)$, which (by (8)) for agent $j = j(\theta^2,\theta)$ strictly dominates announcing $m^j = 0$ and thereby inducing $\tilde{a}(\theta^2)$. Hence, there cannot be an equilibrium in which the outcome is $\tilde{a}(\theta^2)$ with $h(\tilde{a}(\theta^2),\theta) \notin f(\theta)$. We conclude that all equilibrium outcomes are in $f(\theta)$ (after renegotiation).

Q.E.D.

In Theorem 3, hypothesis IV is the assumption that preference reversal exists, whereas $\tilde{e}$ in hypothesis VI is an alternative that is sufficiently bad to ensure that agents can be punished.[13] Hypothesis IV -- which is obviously

---

[13]Hypothesis VI is a good deal stronger than required. The same is true of hypothesis V.

necessary for implementability -- did not have to be assumed explicitly in Theorem 2 because, with a linear Pareto frontier, it was implied by the ability to punish.[14]

We now apply Theorem 3 to our earlier buyer-seller example. To obtain a nonlinear frontier, assume that the buyer and seller are risk-averse with strictly concave von Neumann-Morgenstern utility functions U_b and U_s, but otherwise keep the example the same as before. (Thus, for instance, the buyer's utility in state θ if he trades at price $p(\theta)$ is $U_b(\bar{v} - p(\theta))$.)

One way to induce the seller to make the investment would be to implement the trading price rule $p(\theta) = \bar{v}$, $p(\phi) = \underline{v}$ (since the difference is greater than the seller's cost c). Using Theorem 3, we now show that such a rule can be implemented, thanks to the fact that the buyer and seller are risk-averse. Set $\tilde{a}(\theta)$ = "trade at price $\bar{v}$" and $\tilde{a}(\phi)$ = "trade at price $\underline{v}$". Clearly, hypothesis III of the theorem is satisfied ($a(\theta)$ and $a(\phi)$ are both

---

[14]Theorem 3 is close to the "renegotiation translation" of Theorem 3 in Moore and Repullo [1988]. The mechanism exhibited here, however, is rather more complicated than that in Moore and Repullo. Also, our proof is somewhat involved. The reason for the additional complexity is that we have attempted to deal with possible mixed strategy equilibria, whereas Moore and Repullo did not do so in their Theorem 3.

Theorem 3 is a generalization of Theorem 3 in Maskin and Tirole [1999], which draws a similar conclusion in the particular context of a contracting setting similar to our buyer-seller example.

efficient in both states.) Set $\tilde{b}(\theta,\phi)$ = "trade at price $\frac{1}{4}(\bar{v} + \underline{v})$". $\tilde{b}(\theta,\phi)$ is efficient in both states. And set $\tilde{c}(\theta,\phi)$ = "no trade (and no payment)". $\tilde{c}(\theta,\phi)$ is inefficient in both states. Given that h is determined through Nash bargaining, $h(\tilde{c}(\theta,\phi),\theta)$ = "trade at price $\frac{1}{2}\bar{v}$", and $h(\tilde{c}(\theta,\phi),\phi)$ = "trade at price $\frac{1}{2}\underline{v}$". Notice that the buyer strictly prefers $h(\tilde{b}(\theta,\phi),\theta)$ to $h(\tilde{c}(\theta,\phi),\theta)$ in state θ, and strictly prefers $h(\tilde{c}(\theta,\phi),\phi)$ to $h(\tilde{b}(\theta,\phi),\phi)$ in state ϕ -- thus one half of hypothesis IV of the theorem is satisfied. The other half (reversing the roles of θ and ϕ) is satisfied if we set $\tilde{b}(\phi,\theta) = \tilde{c}(\theta,\phi)$ and $\tilde{c}(\phi,\theta) = \tilde{b}(\theta,\phi)$. Hypothesis V is clearly satisfied, since the price is not bounded.[15] Finally, set $\tilde{e}$ = "trade at price $\tilde{p}$" where $\tilde{p}$ is a payment sufficiently random -- hence sufficiently unattractive to both the buyer and the seller -- that hypothesis VI of the theorem is satisfied.

In fact, there is a simple mechanism that implements the trading price rule $p(\theta) = \bar{v}$, $p(\phi) = \underline{v}$. The seller announces the state. If she announces ϕ, the outcome is "trade at price $\underline{v}$". If she announces θ, then the buyer can either agree or challenge. If he agrees, the outcome is "trade at price $\bar{v}$". If he challenges, the outcome is "no trade, but the buyer pays the seller $\tilde{\ell}$", where $\tilde{\ell}$ is a random payment chosen to satisfy the three inequalities:

---

[15]Strictly speaking, in order to satisfy our initial asumption that A is finite, we ought to restrict the set of feasible prices to lie on a finite grid. But the finiteness assumption is inessential.

$EU_b(\frac{1}{2}\bar{v} - \tilde{\ell}) < U_b(0)$, $EU_b(\frac{1}{2}\underline{v} - \tilde{\ell}) > U_b(\underline{v} - \bar{v})$, and $EU_s(\frac{1}{2}\underline{v} + \tilde{\ell}) < U_s(\underline{v})$.[16] The idea is that in state θ the buyer will agree (he gets $U_b(0)$ by agreeing, but, following the Nash bargain, he gets only $EU_b(\frac{1}{2}\bar{v} - \tilde{\ell})$ from challenging); whereas in state ϕ he will challenge (he only gets $U_b(\underline{v} - \bar{v})$ by agreeing, but, following the Nash bargain, he gets $EU_b(\frac{1}{2}\underline{v} - \tilde{\ell})$ from challenging). The seller has an incentive to announce θ only if she anticipates that the buyer will agree (so that she gets $U_s(\bar{v})$). In state ϕ, after the buyer's challenge and the Nash bargain, the seller only gets $EU_s(\frac{1}{2}\underline{v} + \tilde{\ell})$, rather than the $U_s(\underline{v})$ she gets by announcing ϕ. Unfortunately, the mechanism used to prove Theorem 3 has to be much more intricate than this because the theorem deals with general environments.

The nature and timing of the lottery $\tilde{\ell}$ are critical. If the seller announces θ and the buyer challenges, the randomization is performed mechanically and <u>instantaneously</u>, so there is no time to renegotiate.[17]

The reason the trading price rule $p(\theta) = \bar{v}$, $p(\phi) = \underline{v}$ cannot be

---

[16] Such a lottery is not hard to find. For example, with constant absolute risk aversion, $\tilde{\ell}$ can be chosen to cost the buyer a certainty equivalent of $\frac{3}{4}\bar{v} - \frac{1}{4}\underline{v}$. The first two inequalities are then automatically satisfied; the third will be satisfied for risky enough lottery $\tilde{\ell}$.

[17] It should be noted that the authors disagree with each other about the practicability of such a randomization, see Maskin and Tirole [1999] and Hart and Moore [1999].

implemented in the risk neutral (linear frontier) case has nothing to do with the absence of preference reversal. Our choices of $\tilde{a}(\cdot)$, $\tilde{b}(\cdot,\cdot)$,and $\tilde{c}(\cdot,\cdot)$ satisfy hypotheses III and IV of Theorem 3 both with risk-neutrality and with risk-aversion; that is, preference reversal is guaranteed. And so the crux of the matter is whether it is possible to punish both agents for a deviation (such as announcing different states). In the risk-neutral case, this was impossible, but, as we have seen, it becomes feasible once there is risk aversion.

5. More than Two Agents

We turn now to environments with three or more agents. In implementation theory, there is a typically a significant difference between the cases $n = 2$ and $n > 2$. With only two agents, it may be difficult to determine from a non-equilibrium profile of strategies which agent has deviated. Thus, the most effective punishment in such cases may be to punish both agents. The outcome $\tilde{e}$ played the role of such a mutual punishment in Theorem 3. And it was the unavailability of this sort of "bad" outcome that made renegotiation so constraining in Theorem 2.

Once there are at least three agents, by contrast, a unilateral deviator from equilibrium can be detected by comparing his behavior with that of the other agents. Thus we can dispense with the requirement of a bad outcome $\tilde{e}$: in Theorem 4, we drop hypothesis VI from Theorem 3.

Theorem 4 in fact shows that, for $n > 2$, the first two hypotheses from Theorem 3 -- hypotheses III and IV -- are necessary conditions for

implementability. And, together with hypotheses VII and VIII below, they are sufficient. The "gap" between the necessary and sufficient conditions is very small: hypotheses VII and VIII are extremely mild and will be automatically satisfied in almost all applications.[18]

Theorem 4: Assume that $n \geq 3$. The SCR f can be implemented in subgame-perfect equilibrium with renegotiation function h only if hypotheses III and IV of Theorem 3 are satisfied. Conversely, f can be implemented in subgame-perfect equilibrium with renegotiation function h if, in addition to hypotheses III and IV, it is the case that, in any state $\theta \in \Theta$:

(VII) no alternative in $\{h(\tilde{a}(\xi),\theta) | \xi \in \Theta\}$ is maximal in ΔA for any agent; and

(VIII) no alternative in ΔA is maximal for two or more agents.

---

[18]Hypothesis VII is a weakened form of hypothesis V from Theorem 3. Hypothesis VIII ensures that the "no veto power" condition from standard implementation theory (see footnote 10) holds vacuously.

In fact, hypotheses VII and VIII are a good deal stronger than required for the sufficiency result in Theorem 4. We could close the gap between between the necessary and sufficient conditions, but only at the cost of considerably complicating the statement of the theorem (cf. Abreu and Sen (1990) and footnote 24 of Moore and Repullo (1988), on subgame perfect implementability without renegotiation).

Proof: To establish necessity, suppose, contrary to the theorem, that, for all $\tilde{a}(\cdot)\colon \Theta \longrightarrow \Delta A$ for which

$$h(\tilde{a}(\xi),\xi) \in f(\xi) \text{ for all } \xi \in \Theta, \tag{9}$$

there exist $\theta,\phi \in \Theta$ such that $h(\tilde{a}(\theta),\phi) \notin f(\phi)$ and, for all i,

$$h(\tilde{b},\theta)R_i(\theta)h(\tilde{c},\theta) \quad \text{iff} \quad h(\tilde{b},\phi)R_i(\phi)h(\tilde{c},\phi) \quad \text{for all } \tilde{b},\tilde{c} \in \Delta A. \tag{10}$$

Now, if f is implementable in subgame-perfect equilibrium with renegotiation function h, there exist a game form g and function $\tilde{a}(\cdot)\colon \Theta \longrightarrow \Delta A$ satisfying (9) such that, for all ξ, $\tilde{a}(\xi)$ is a subgame-perfect equilibrium outcome in state ξ. But then there exist θ and ϕ, with $h(\tilde{a}(\theta),\phi) \notin f(\phi)$, such that $\tilde{a}(\theta)$ is a subgame-perfect equilibrium outcome in state θ and, from (10), $\tilde{a}(\theta)$ is also a subgame-perfect equilibrium in state ϕ. But because $h(\tilde{a}(\theta),\phi) \notin f(\phi)$, this contradicts the assumption that g implements f with renegotiation function h. We conclude that hypotheses III and IV are necessary.

As for sufficiency, let us start with some definitions. Given that the lottery $h(\tilde{b}(\theta,\phi),\theta)$ may not be Pareto optimal in state θ, define $\hat{h}(\tilde{b}(\theta,\phi),\theta)$ to be the (Pareto optimal) outcome that is reached if renegotiation occurs <u>before</u> the resolution of the randomness in $\tilde{b}(\theta,\phi)$. Because $h(\tilde{a}(\theta),\theta)$ is also Pareto optimal (from hypothesis III), there exists, for all ϕ such that $h(\tilde{a}(\theta),\phi) \notin f(\phi)$, an agent $j(\theta,\phi)$ such that

$$h(\tilde{a}(\theta),\theta)R_{j(\theta,\phi)}(\theta)\hat{h}(\tilde{b}(\theta,\phi),\theta). \tag{11}$$

We can now construct a stage mechanism that implements f in subgame-perfect equilibrium with renegotiation function h. In the first stage each agent i announces a state, θ^i, and a nonnegative integer m^i. If there exist $\theta' \in \Theta$ and j such that $(\theta^i, m^i) = (\theta', 0)$ for all $i \neq j$, then the outcome of the mechanism is $\tilde{a}(\theta')$ -- unless $h(\tilde{a}(\theta'), \theta^j) \notin f(\theta^j)$, $j = j(\theta', \theta^j)$, and $m^j > 0$, in which case the mechanism moves to stage $2(\theta', \theta^j)$. In all other cases in this first stage, the agent i choosing the highest integer m^i (with ties broken by a coin flip) gets to choose any alternative in ΔA.

In stage $2(\theta', \theta^j)$, each agent i announces a nonnegative integer n^i. If $n^i \neq 0$ for at most one agent i, then agent $j(\theta', \theta^j)$ gets to choose any alternative in ΔA. If $n^i \neq 1$ for at most one agent i, then the outcome is $\tilde{b}(\theta', \theta^j)$ unless $n^{k(\theta', \theta^j)} > 1$, in which case the outcome is $\tilde{c}(\theta', \theta^j)$, where these outcomes and $k(\theta', \theta^j)$ are defined in hypothesis IV. In all other cases, the agent i setting the highest integer n^i (with ties broken by a coin flip) gets to choose any alternative in ΔA.

Let us verify that this mechanism implements f with renegotiation function h. Suppose throughout that θ is the true state.

We claim that the following is a subgame-perfect equilibrium, whose outcome, following renegotiation, is $h(\tilde{a}(\theta), \theta)$ -- which, by hypothesis III, is contained in $f(\theta)$. In the first stage, each agent i sets $(\theta^i, m^i) = (\theta, 0)$. If play reaches any stage $2(\theta, \theta^j)$, then each agent i sets $n^i = 1$. If play reaches any stage $2(\theta', \theta^j)$ where $\theta' \neq \theta$, then each agent i sets $n^i = 0$. Finally, in any subgame where an agent gets to choose an alternative in ΔA, he chooses his favorite alternative.

The claim is straightforward to verify. The heart of the matter is that if $h(\tilde{a}(\theta),\theta^j) \notin f(\theta^j)$ for some θ^j, then, by (11), agent $j(\theta,\theta^j)$ has no incentive to announce $m^{j(\theta,\theta^j)} > 0$ at the first stage. And if play reaches stage $2(\theta,\theta^j)$, agent $k(\theta,\theta^j)$ has no incentive to announce $n^{k(\theta,\theta^j)} > 1$, since, by (4), he weakly prefers $\tilde{b}(\theta,\theta^j)$ to $\tilde{c}(\theta,\theta^j)$.

It remains to show that all other equilibrium outcomes are also in $f(\theta)$.

Start at stage $2(\theta',\theta)$. Suppose that in equilibrium there is a positive probability that $n^i > 0$ for some agent i. Then at least one of the other agents $k \neq i$ has a strict incentive to announce a suitably large n^k. To see why, consider the three possible consequences of such an announcement. First, agent k may not affect the outcome (either because the other agents are all announcing zero, or because the other agents are all announcing 1 and $k \neq k(\theta',\theta)$). Second, agent k may change the outcome from $\tilde{b}(\theta',\theta)$ to $\tilde{c}(\theta',\theta)$ (because $k = k(\theta',\theta)$ and the other agents are all announcing 1). Third, with arbitrarily high probability, agent k may get to implement his favorite alternative in ΔA (because he wins the "integer game" $\{n^1,\ldots,n^n\}$). Given that there is a positive probability that $n^i > 0$, it therefore follows from (5) and hypothesis VIII that, for at least one $k \neq i$, agent k has a strict incentive to announce a suitably large n^k. Now if this agent k always announces $n^k > 1$ then, by hypothesis VIII, there is some other agent who strictly prefers to announce a yet larger integer; and clearly there can be no equilibrium of this kind. Thus, the only continuation equilibrium at stage $2(\theta',\theta)$ has $n^i = 0$ for all i. That is, $j(\theta',\theta)$ chooses his favorite alternative in ΔA.

Now return to the first stage. Suppose, with positive probability,

either some agent i announces $m^i > 0$, or two agents announce different states. Then one of the agents, j say, has a strict incentive to announce $\theta^j = \theta$ together with a suitably large m^j. To see why, consider the two possible consequences of such an announcement. First, agent j may not affect the outcome (because the other agents are announcing zero and a common θ', and $j \neq j(\theta',\theta)$). Second, agent j will get to implement his favorite alternative (either because $j = j(\theta',\theta)$, the other agents are all announcing θ' and zero, and, as we have just shown, $j(\theta',\theta)$ gets to choose his favorite alternative in the unique continuation equilibrium; or because, with arbitrarily high probability, agent j wins the "integer game" $\{m^1,\ldots,m^n\}$). Given that there is a positive probability that either some agent i announces $m^i > 0$ or two agents announce different states, it therefore follows from hypothesis VIII that at least one agent j has a strict incentive to announce $\theta^j = \theta$ together with a suitably large m^j. Now if this agent j always announces $n^j > 0$ then, by hypothesis VIII, there is some other agent who strictly prefers to announce a yet larger integer; and clearly there can be no equilibrium of this kind.

Hence the only candidate equilibrium at the first stage has $(\theta^i,m^i) = (\theta',0)$ for all i, with outcome $\tilde{a}(\theta')$. However, if $h(\tilde{a}(\theta'),\theta) \notin f(\theta)$, then this cannot be an equilibrium either, because, by hypothesis VII, agent $j(\theta',\theta)$ strictly prefers to announce $(\theta,1)$ so as to move the mechanism on to stage $2(\theta',\theta)$ and allow him to choose his favorite alternative.

Q.E.D.

Let us apply Theorem 4 to our buyer-seller example. Although there are only two agents in that model, we can introduce a passive third agent as a

way of breaking the "balanced budget" constraint. Without a third agent, the seller must receive whatever the buyer pays. (The agents cannot agree to throw money away because such an agreement would be renegotiated.) But with a third party, the equality between the buyer's payment and the seller's receipt need not hold: the third party could get some of the money himself. Indeed, the presence of a third party allows the buyer and the seller to be jointly punished when there is a deviation from equilibrium; for certain configurations of strategies, they might both have to pay him something.[19] And it does not matter if the third party doesn't observe the state.

More precisely, consider the simple mechanism we proposed at the end of Section 3 for inducing the seller to make her investment when the agents are risk averse. Let us now assume that the agents are risk neutral, but that there is a third party who can act as a financial sink. (The mechanism works equally well if the agents are risk averse.) Keep the mechanism the same except that, following an announcement of θ by the seller, if the buyer challenges then the outcome is "no trade, but the buyer pays the third party $\frac{3}{4}\bar{v} - \frac{1}{4}\underline{v}$ (there are no payments to or from the seller)". As before, the idea is that in state θ the buyer will agree (he gets zero by agreeing, but, following the Nash bargain, he gets only $-\frac{1}{4}(\bar{v} - \underline{v})$ from challenging); whereas in state ϕ he will challenge (he only gets $-(\bar{v} - \underline{v})$ by agreeing, but, following the Nash bargain, he gets $-\frac{3}{4}(\bar{v} - \underline{v})$ from challenging). The

---

[19]The introduction of a third party might, however, create other problems, notably the possibility of collusion (see Hart and Moore [1988, 1999]). It is a matter of debate whether such collusion can be ruled out contractually. But this issue is beyond the scope of this paper.

seller has an incentive to announce θ only if she anticipates that the buyer will agree (so that she gets $\bar{v}$). In state ϕ, after the buyer's challenge and the Nash bargain, the seller only gets $\frac{1}{2}\underline{v}$, rather than the $\underline{v}$ she gets by announcing ϕ.

We introduced the concept of renegotiation-monotonicity in Section 2 as the natural translation of ordinary monotonicity into a setting where renegotiation can occur. For completeness, let us formally state the result that invokes it.

Theorem 5: Assume that $n \geq 3$. The SCR f can be implemented in Nash equilibrium with renegotiation function h only if f and h satisfy renegotiation-monotonicity. Conversely, if f and h satisfy renegotiation-monotonicity and, in all states $\theta \in \Theta$, hypothesis VIII of Theorem 4 holds, then f can be implemented in Nash equilibrium with renegotiation function h.

Proof: A straightforward translation of the proof in Maskin [1999]. (Maskin [1999] invokes no veto power, which, as we have already noted, is satisfied vacuously when hypothesis VIII holds.)

REFERENCES

Abreu, D. and A. Sen (1990), "Subgame Perfect Implementation: A Necessary and Almost Sufficient Condition," Journal of Economic Theory,50, 285-299.

Aghion, P., M. Dewatripont, and P. Rey (1994), "Renegotiation Design with Unverifiable Information," Econometrica, 62, 257-282.

Chung, T. (1992), "Incomplete Contracts, Specific Investments, and Risk Sharing," Review of Economic Studies, 58, 1031-1042.

Corchon, L. (1996), The Theory of Implementation of Socially Optimal Decisions in Economics, New York: St. Martin's Press.

Green, J. and J.J. Laffont (1988), "Contract Negotiation and the Underinvestment Effect," mimeo.

Hart, O. and J. Moore (1988), "Incomplete Contracts and Renegotiation," Econometrica, 56, 755-785.

Hart, O. and J. Moore (1999), "Foundations of Incomplete Contracts," Review of Economic Studies, forthcoming.

Hermalin, B. and M. Katz (1991), "Moral Hazard and Verifiability: The Effects of Renegotiation in Agency," Econometrica, 59, 1735-1754.

Hurwicz, L., E. Maskin, and A. Postlewaite (1995), "Feasible Nash Implementation of Social Choice Rules where the Designer Does Not Know Endowments or Production Sets," in J. Ledyard (ed.), The Economics of Informational Decentralization, Kluwer Academic Publishers.

Maskin, E. (1999), "Nash Equilibrium and Welfare Optimality," Review of Economic Studies, forthcoming.

Maskin, E. and J. Tirole (1999), "Unforeseen Contingencies and Incomplete Contracts," Review of Economic Studies, forthcoming.

Moore, J. and R. Repullo (1988), "Subgame Perfect Implementation," Econometrica, 56, 1191-1220.

Moore, J. (1992), "Implementation, Contracts, and Renegotiation in Environments with Complete Information," in J.J. Laffont (ed.), Advances in Economic Theory, Cambridge: Cambridge University Press.

Noldecke, G. and K. Schmidt (1995), "Option Contracts and Renegotiation: A Solution to the Hold-Up Problem," Rand Journal of Economics, 26, 163-179.

Osborne,M. and A. Rubinstein (1994), A Course in Game Theory, Cambridge, MA: MIT Press.

Palfrey, T. (1992), "Implementation in Bayesian Equilibrium: The Multiple Equilibrium Problem in Mechanism Design," in J.J. Laffont (ed.), Advances in Economic Theory, Cambridge: Cambridge University Press.

Palfrey, T. (1998), "Implementation Theory," forthcoming in R. Aumann and S. Hart (eds.), Handbook of Game Theory, Vol. 3, Amsterdam: North Holland.

Rubinstein, A. and A. Wolinsky (1992), "Renegotiation-Proof Implementation and Time Preferences," American Economic Review, 600-614.

Segal, I. (1999), "Complexity and Renegotiation: A Foundation for Incomplete Contracts," Review of Economic Studies, forthcoming.