Efficient and Durable Decision Rules: A Reformulation

Vincent P. Crawford

Econometrica, Volume 53, Issue 4 (Jul., 1985), 817-836.

Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198507%2953%3A4%3C817%3AEADDRA%3E2.0.CO%3B2-V

Econometrica, Vol. 53, No. 4 (July, 1985)

EFFICIENT AND DURABLE DECISION RULES: A REFORMULATION[1]

BY VINCENT P. CRAWFORD

This paper studies the limits of contracting as a method for achieving efficient allocation, with particular attention to how informational asymmetries interact with the timing of commitment to a mechanism. There are arguments to suggest, in the spirit of the Coase "Theorem," that if agents can agree on a mechanism before observing their private information (or, a fortiori, if information is perfect or symmetric), they can realize an incentive-efficient allocation. If, however, agents observe their private information before contracting, there may be further restrictions, due to information leakage during the process of bargaining over mechanisms, on what they can achieve by contract. These restrictions are characterized and compared to those proposed for this setting by Holmstrom and Myerson [6]. It is also shown that there is at least one specification of the rules that govern mechanism design that makes it possible for agents to achieve, contracting after they observe their private information, the same incentive-efficient allocations that are attainable when they can commit themselves to a mechanism before observing their private information.

1. INTRODUCTION

WHAT ARE THE LIMITS of contracting as a way to ensure efficient resource allocation? When agents have perfect information, this basic question can be given a simple answer. Suppose that agents must play a game, but that before play they have an opportunity to make complete, perfectly enforceable agreements about how to play the game. Once the rules for making such agreements are specified, the contracting process together with the underlying game can be analyzed as a non-cooperative game. For reasonable specifications of the rules of contracting, there are arguments to suggest, in the spirit of the Coase "Theorem," that rational agents can reach an agreement that is individually rational (i.e., better for each agent than no agreement) and Pareto-efficient.

This answer is readily extended to the case where there is uncertainty, but agents have symmetric information. Here, individual rationality and efficiency are taken to refer to agents' preferences over agreements before the uncertainty is resolved, and efficiency is defined relative to the limitations, if any, on agents' ability to make complete contingent contracts. This yields an exact formal analogy between the perfect- and symmetric-information cases, so that the conclusion that agents can reach efficient, individually rational agreements remains valid, properly reinterpreted.

[1] This paper is an extensively revised version of Harvard Institute of Economic Research Discussion Paper 962. I am grateful to the Harvard Economics Department for providing a congenial research environment, and to the Harvard Divinity School, where the idea for the mechanism-design process studied in Section 4 first came to me. I have benefited from the suggestions of Peter Cramton, Jerry Green, Theodore Groves, Bengt Holmstrom, Ehud Kalai, Co-Editor James Mirrlees, Roger Myerson, Michael Rothschild, Joel Sobel, and an anonymous referee; and from the comments of seminar audiences in the Harvard-MIT Game Theory Workshop, the UCSD Theory Workshop, the Minnesota IMA Workshop on Bayesian Analysis in Economics and Game Theory, and the Stanford IMSSS Workshop. The National Science Foundation provided partial financial support through Grants SES-8106912 to Harvard and SES-8204038 to UCSD.

Significant new difficulties arise when information is asymmetric. Incentive schemes, commonly called mechanisms, are the natural objects of choice in this context. Thus, it seems natural to evaluate the limits of contracting by assuming that agents can make agreements about the mechanism that will be used to govern play in the underlying game, and that those agreements are evaluated according to rational expectations of the effects of the incentives they create. Two cases can be distinguished. In the first, agents can commit themselves to an agreement before observing their private information. This case becomes equivalent to the perfect- and symmetric-information cases when feasibility and efficiency are redefined, as in Myerson [10], to reflect incentive constraints. In the language of Holmstrom and Myerson [6] (henceforth "HM") it can be argued that agents will agree on an ex ante incentive-efficient mechanism that is individually rational in the sense appropriate to the situation. This is the natural generalization of the perfect- and symmetric-information conclusions.

The second case, where agents cannot make binding agreements until after they have observed their private information, is probably more realistic and certainly more difficult to analyze. The difficulty is due to the fact that endogenous mechanism design must, to yield good results, aggregate agents' preferences over mechanisms. These preferences are influenced by agents' private information, which may therefore be "leaked" in this case in the process of reaching an agreement, altering the incentives created by the agreed-upon mechanism. Even if the effects of this information leakage are rationally anticipated, they can prevent agents from bringing about some desirable outcomes that would be attainable if they could make binding agreements before observing their private information. This difficulty, and the associated welfare effects, are customarily avoided in the incentives literature by assuming either that a mechanism is imposed exogenously; that agents choose a mechanism before observing their private information; or, as in the principal-agent literature (see Myerson [11] for a notable exception), that no agent who plays an active role in mechanism design has any private information.

HM [6] extend the techniques developed in Myerson's [11] analysis of the "informed-principal" problem to describe this information-leakage problem and propose a characterization of the limits of contracting that stem from it. They call the outcomes that rational agents might agree on despite information-leakage problems *durable*; durability can be defined loosely as the ability to withstand alternative proposals when a unanimous vote is required for adoption. The present paper has three purposes. First, it supplements HM's analysis by providing a fuller discussion of some of the issues that arise in modeling endogenous mechanism design with information-leakage problems. Second, it proposes and defends an alternative to HM's characterization of durability. The alternative is in the spirit of noncooperative bargaining theory, and has the important advantage that it allows a complete model of agents' mechanism-proposal decisions; this seems essential for a full understanding of the effects of information leakage.

Finally, the paper argues that for game-theoretic analysis of endogenous mechanism design to be possible, the rules that govern it must, at some level, be

imposed by the modeler rather than derived within the model. These exogenous rules can usefully be viewed as the ultimate source of the restrictions due to information leakage. It is shown that one reasonable specification of the rules renders unattainable some desirable outcomes agents could achieve by committing themselves to a mechanism before observing their private information; these restrictions are analogous to, but generally different from, HM's durability restrictions. Another simple specification has the remarkable effect of entirely eliminating such restrictions, allowing agents to achieve the same incentive-efficient outcomes that would be attainable if they could design a mechanism behind the "veil of ignorance." These rules accomplish this by making it possible for agents endogenously to restrict the language they use in bargaining over mechanisms, in a way that prevents the encoding of private information in mechanism proposals but leaves the language rich enough to permit efficient aggregation of agents' preferences over mechanisms. The strikingly different implications of these two simple specifications suggest that it would be useful to learn more about the general characteristics of the social institutions that govern endogenous mechanism design in practice.

The paper is organized as follows. Section 2 begins with a description of the economic environment and a glossary of the terms and concepts used here and in [6]. The section concludes with a discussion of HM's [6, Section 6] main example and a description of their unanimous-voting characterization of durability. Section 3 presents an alternative characterization in the spirit of noncooperative bargaining theory, in which agents bargain over mechanisms, but the rules of bargaining are exogenous. Its implications in HM's example are discussed. Section 4 describes a simple modification of the rules of bargaining over mechanisms assumed in Section 3 that has the effect of eliminating "durability" restrictions. Section 5 is the conclusion.

2. THE MODEL AND TERMINOLOGY

This section describes the economic environment and the terminology that will be used. Both are the same as in [6], except where noted. There are n agents, indexed $i = 1, \ldots, n$, who must play a game. The game consists of a specification of what strategies are feasible for each agent, the mapping from agents' strategy choices to physical outcomes, and agents' information and preferences. In what follows, I shall refer to this game as the *underlying game* to distinguish it from the rules of bargaining over mechanisms; and I shall assume that if agents do not reach agreement on a mechanism, play in the underlying game is governed by a particular (Bayesian Nash) equilibrium, predictable in advance by all agents.

Agents begin with (possibly different) prior probability distributions over the state of the world; all agents agree, however, on which states have zero prior probability. Agents have risk preferences representable by von Neumann-Morgenstern utility functions, defined over physical outcomes and states. Agent i has a finite set of T_i possible *types* indexed $t_i = 1, \ldots, T_i$; t_i can be thought of as a parameter of his preferences and beliefs, which includes all of his private informa-

tion. The vector $t = (t_1, \ldots, t_n)$ is therefore a complete description of the state of the world; all other features of the economy are common knowledge.

The term *symmetric* (equivalently, *complete*) information refers to situations where agents have exactly the same information (as common knowledge) when decisions are made. With *asymmetric* (equivalently, *incomplete*) information, some decisions must be made when agents have different information. Information is asymmetric, in terms of the notation just introduced, unless $T_i = 1$ for all $i = 1, \ldots, n$.

A *decision rule* is a functional relationship between outcomes (defined to include probability distributions) and agents' observed types—in other words, a type-contingent allocation. A *mechanism* is a set of physical rules that modify the rules of the underlying game. A *direct-revelation* mechanism, for example, asks agents to report their types and selects a joint strategy choice in the underlying game as a function of their reports. Thus, a mechanism as defined here, together with agents' responses to it, determines a decision rule. HM use the terms decision rule and mechanism interchangeably, but the distinction preserved by my more traditional use of the latter term is important here.

An *incentive-compatible* mechanism is a direct-revelation mechanism in which truthful reporting is an equilibrium. The interest of incentive-compatible mechanisms is enhanced by the "Revelation Principle" argument (see, for example, [**6**, Section 3]) that any decision rule that can be realized as an equilibrium of some (exogenously imposed) mechanism, can also be realized in the truthful equilibrium of an incentive-compatible mechanism. Thus, if it is assumed that coordination on any desired equilibrium can be achieved, there is no loss of generality, when mechanisms are imposed exogenously or designed before agents observe their private information, in restricting attention to incentive-compatible mechanisms.

When information is asymmetric, the standard physical notion of feasibility must be refined by the requirement that agents have incentives to reveal, directly or indirectly, the information on which it is desired to make the outcome depend. *Incentive-efficiency* is defined in the same way as ordinary Pareto-efficiency, but reflects the resulting incentive constraints (see Myerson [**10**] for a development of this idea in the context of bargaining). The terms *ex ante*, *interim*, and *ex post* refer, respectively, to before agents have observed their types; after they have observed their own types, but not other agents' types; and after all types have been revealed. Thus, given the Revelation Principle, an *interim incentive-efficient* mechanism can be defined as an incentive-compatible mechanism such that there is no other incentive-compatible mechanism that is at least as good, in the truthful equilibrium, for every type of every agent, and strictly better for at least one type of one agent. Similarly, an interim incentive-efficient decision rule is one associated with the truthful equilibrium of some interim incentive-efficient mechanism. Other terms are defined analogously; see [**6**, Sections 3 and 4] for formal definitions.

A mechanism or decision rule is said to be *individually rational* if it is weakly preferred, by every agent, to the consequences of nonparticipation, which are

taken here to be noncooperative play of the underlying game. The bargaining approach used here requires more attention to individual rationality than HM's unanimous-voting approach, because they take the initial mechanism as an exogenous status quo and do not allow agents not to participate in this mechanism. I shall also simplify the treatment of individual rationality, by assuming that all feasible mechanisms allow each agent to choose unilaterally not to participate, and that agents cannot commit themselves ex ante to participate in the mechanism or to report truthfully at the interim stage. The former assumption has the effect of subsuming individual rationality in incentive-compatibility; see Myerson [**11**, Section 2] for a similar treatment. The latter assumption makes ex ante and interim incentive-compatibility and individual rationality, which differ in general, equivalent; HM [**6**, Sections 3 and 4] also maintain this assumption.

HM [**6**, Section 6] present an example to illustrate the problems caused by information leakage when mechanism design takes place at the interim stage, and to motivate their notion of durability. The example has two agents, 1 and 2, each of whom has two possible types, a and b; all four possible combinations are equally likely. There are three feasible "pure" decisions, A, B, and C; randomization over these decisions is also allowed. Participation in the mechanism that is selected is required. Finally, agents' von Neumann-Morgenstern utilities for the various decisions, which are given in the following table (reproduced from [**6**, p. 1809]), depend only on their own types, so that informational interactions are confined to the incentive constraints.

	u_{1a}	u_{1b}	u_{2a}	u_{2b}
$d = A$	2	0	2	2
$d = B$	1	4	1	1
$d = C$	0	9	0	−8

In the example, the direct-revelation mechanism that selects decision B whenever 2 reports b and, otherwise, A when 1 reports a and C when 1 reports b, is incentive-compatible and uniquely maximizes the sum of agents' ex ante expected utilities in the set of incentive-compatible mechanisms. It is therefore ex ante and interim incentive-efficient, so that (as HM's Theorem 1 shows) it cannot be common knowledge that some other mechanism is better for all four types. However, it is common knowledge that the alternative mechanism "A-for-sure" is better for both types of agent 2 and type a of agent 1. This raises three closely related issues:

1. After observing their types, agents may unanimously prefer a given incentive-compatible mechanism that is *not* interim incentive-efficient to one that is. (The example does not show this directly, because its simple structure renders the mechanisms in question interim incentive-efficient. But it makes clear that richer examples with the desired property are possible.) The example also shows that it may be possible for an agent to discern this unanimity of preferences from his private information. Thus, agents' preferences at the interim stage do not favor

the design of interim incentive-efficient mechanisms as strongly as their ex ante preferences favor ex ante incentive-efficiency.

2. The fact that agents cannot commit themselves not to recontract at the interim stage may prevent them from implementing some desirable incentive-compatible mechanisms.

3. The rules for proposing and adopting mechanisms influence what agents can achieve in mechanism design at the interim stage, through their effect on the inferences agents draw from other agents' mechanism proposals (or failures to make proposals). In HM's example, if agent 1 observes type a, he will take any opportunity to propose A-for-sure, which he knows is the best possible mechanism for himself and agent 2. Failure to make this proposal will therefore convince 2 that 1 has observed b, which destroys the incentive-compatibility of the original ex ante incentive-efficient mechanism.

HM present a characterization, motivated by these three issues, of mechanisms that rational agents might agree upon at the interim stage. They call such mechanisms *durable*, and define them as those that are both incentive-compatible and able to withstand all other mechanisms when a unanimous vote, with the alternative proposed anonymously, is required to replace the status quo mechanism. Durability requires, more precisely, that there be a nontrivial (see [6, Section 7]) equilibrium of the voting game in which the status quo mechanism gets at least one vote. In applying this definition, HM assume that agents expect the vote in question to determine the final choice of mechanism. Voting is simultaneous, with agents rationally conditioning their votes on the information revealed by the result and the effects of that information leakage on the performance of the mechanisms being considered. Finally, HM do not describe how the original mechanism became the status quo. This requires one of two interpretations of their characterization: either the original mechanism became the status quo without any information leakage—perhaps by being imposed exogenously or proposed by an agent with no private information—or agents' priors already include the effects of past information leakage.

HM's characterization is simple and frictionless, and gives the answer suggested by the Coase "Theorem" when information is perfect or symmetric, or, suitably reinterpreted, when information is asymmetric but agents choose a mechanism at the ex ante stage. Studying their characterization when mechanism design takes place at the interim stage helps to clarify issue 2 above, and sheds some light on issues 1 and 3. But because their analysis is not based on a complete model of the process by which agents propose and adopt mechanisms, it leaves room for further clarification of all three issues. The next section presents an alternative characterization of the possible outcomes of rational endogenous mechanism design at the interim stage, in the spirit of noncooperative bargaining theory. Because bargaining theory suggests natural models of agents' mechanism proposals and how the mechanism-design process aggregates their preferences, the alternative characterization has a significant advantage in understanding issues 1 and 3. To make this approach tractable, however, I specify the rules of bargaining over mechanisms in a way that trivializes issue 2.

3. A REFORMULATION OF DURABILITY

This section describes and motivates an alternative formulation of HM's concept of durability, based on noncooperative bargaining theory (see Harsanyi [5] or Crawford [3, Section 2] and the references cited there). In this formulation, endogenous mechanism design takes place within a two-stage game. In the first stage, agents bargain over the mechanism that will be used to determine how the underlying game is played; the rules that govern bargaining over mechanisms, described further below, cannot be altered. In the second stage, agents play the underlying game, responding to the incentives created by the agreed-upon mechanism. Agents' first-stage evaluations of mechanisms are based on rational expectations of their second-stage effects, taking into account the incentive effects of any information leakage that occurs while the mechanism is being agreed upon.

Because a mechanism's welfare properties may be altered by the effects of information leakage, it is the decision rules that emerge from the mechanism-design process, which take these effects into account, that are of primary interest, rather than the mechanism that is selected. For my purposes, mechanism-design behavior can be modeled by assuming that agents play strategies that are in *perfect* equilibrium, in the sense that their strategies are in equilibrium in every subgame (of incomplete information) and their beliefs are derived from Bayes' Law whenever it is applicable. (See Kreps and Wilson [9] for a good discussion of further possible refinements; the notion of equilibrium used here is what they call "extended subgame perfectness.") The decision rules that arise in perfect equilibrium are the natural analogs of HM's durable mechanisms; I shall call them *attainable* decision rules to preserve the distinction.

Before describing the specific assumptions about the rules of bargaining over mechanisms maintained in this section, it may be helpful to record two observations about endogenous mechanism design. First, with a complete model of the mechanism-design process, incentive-compatibility is immediately seen to be a necessary condition for attainability, because the entire game can be viewed, via the Revelation Principle, as a direct-revelation mechanism to implement any attainable decision rule. Attainability is, of course, more restrictive than incentive-compatibility, for two main reasons. Because the mechanism-design process is responsive to agents' preferences, attainability, like durability, is more closely related to efficiency than to feasibility: really undesirable outcomes are unlikely to emerge from rational endogenous mechanism design, even when information-leakage problems prevent agents from dealing as effectively as possible with incentive problems. Also, despite this tendency toward incentive-efficiency, the constraint that an attainable decision rule must arise in perfect equilibrium is typically restrictive enough to rule out most incentive-efficient outcomes as well.

The second observation concerns my assumption that the rules of bargaining over mechanisms cannot be altered. It is easy to imagine more general specifications, in which agents can amend the rules, amend the rules for amending the rules, and so on. But it is clear on reflection that unless this sequence comes to a halt at some level, there is no place to anchor a game-theoretic analysis.

Something must be specified exogenously, and it is this specification that ultimately determines the attainability restrictions. Given that exogenous specification cannot be avoided, it seems a sensible research strategy to begin with the simplest rules of bargaining that are rich enough to address the issues of interest, and to assume that they govern mechanism design at the highest level—i.e., that agents are free to bargain over mechanisms, but they cannot amend the rules of bargaining.

These considerations suggest a specification based on Nash's [12] "demand game," suitably generalized to allow agents with asymmetric information to bargain over mechanisms. (Binmore [1] and Harsanyi [5] provide useful discussions of the demand game and the problems encountered in generalizing it to allow asymmetric information.) In the perfect-information demand game studied by Nash, agents make simultaneous demands; each agent's demand specifies a utility level for himself. If these demands are *compatible*, in the sense that they could be realized by some choice of correlated strategies in the underlying game, they are taken as a binding agreement, with each agent receiving exactly his demanded utility level. If not, the bargaining process ends with no agreement, and the underlying game is therefore played noncooperatively.

Nash observed that any pair of demands that yields an individually rational, Pareto-efficient outcome is in equilibrium in the demand game. It is also possible to show under weak assumptions that strongly individually rational, Pareto-efficient pairs are the only pure-strategy, trembling-hand perfect equilibria. (A *strongly* individually rational outcome is one where the implied preference is strict for each agent. Trembling-hand perfectness here rules out trivial equilibria in which each agent makes a demand so high that other agents' demands cannot be reconciled with it and strong individual rationality. Such demands are weakly dominated by demands that have some chance of yielding a beneficial agreement; see [6, Section 7] for an analogous approach to trivial equilibria in HM's voting game. The restriction to pure strategies rules out a continuum of nontrivial but inefficient mixed-strategy equilibria.) Thus, because an agent's demand is a best response to rational predictions of other agents' demands, the simultaneity of demands does not preclude efficient aggregation of agents' preferences. Nash's identification of the efficient and individually rational demand-game equilibria, which he presented as part of a noncooperative rationalization of his axiomatic bargaining solution, can be viewed as a formalization of the Coase Theorem, as discussed in the Introduction. It is interesting not for its familiar conclusion, but because it shows why symmetric information and complete contingent contracts are necessary for that conclusion, indicates how the conclusion must be modified when agents with asymmetric information choose a mechanism at the ex ante stage, and provides a framework in which it is possible to evaluate the limits of contracting at the interim stage in dealing efficiently with incentive problems.

I shall now describe the simplest asymmetric-information generalization of the demand game that allows agents enough flexibility to design incentive-efficient mechanisms at the ex ante stage. For this specification of the rules of bargaining over mechanisms, all ex ante incentive-efficient decision rules are attainable when

mechanism design takes place ex ante, but interim incentive-efficient decision rules are not generally attainable in mechanism design at the interim stage. The specification therefore implies nontrivial attainability restrictions, analogous to HM's durability restrictions. Section 4 presents an alternative, slightly more complex specification of the rules that does make interim incentive-efficient decision rules attainable at the interim stage. This demonstrates the sensitivity of attainability to the rules that govern mechanism design, and shows that it is possible in theory for social institutions to deal efficiently with the information-leakage problems that arise when mechanism design takes place at the interim stage.

For the rules that govern mechanism design to be implementable, they must use only information that is common knowledge to agents. I shall assume, as is customary in the incentives literature, that the rules can make full use of this information; Kalai and Rosenthal [7] have shown that it may be possible for an outside planner to enforce such rules even if he has less information than agents have in common. The key to generalizing the demand game is that, because all private information is summarized by agents' types, type-contingent demands provide a language flexible enough to bargain over mechanisms while respecting the limitations of common knowledge.

Thus, let agents simultaneously and publicly announce demands, where agent i's demand, denoted $u^i = (u^i_1, \ldots, u^i_{T_i})$, is a vector of proposed expected utilities, one for each of his types. If there is some incentive-compatible mechanism that would realize the demands $(u^1, \ldots, u^n)$ exactly in the absence of information leakage, they are said to be *compatible*. Compatible demands are implemented by imposing such a mechanism, whether or not agents draw inferences during the process of bargaining over mechanisms. Incompatible demands cause the mechanism-design process to end with no agreement, so that the underlying game is played noncooperatively.

In implementing these bargaining rules, it is simplest to assume that a normalization of agents' types' von Neuman-Morgenstern utility functions, and a fixed rule for determining which mechanism is imposed when (as is typical) there is more than one with the required property, are publicly announced at the start. The former assumption could be dispensed with, if desired, by expressing demands in physical terms—customarily assumed to have common-knowledge scaling—with the understanding that it is the associated type-contingent utility levels, not the physical actions themselves, that must be implemented. The latter assumption seems harder to dispense with. In particular, because the performance of a mechanism is usually sensitive to agents' information, it might seem desirable to make the choice of mechanism when demands are compatible depend on agents' inferences. But these inferences are both endogenously determined as part of the equilibrium and unobservable. Although allowing agents to bargain over this choice might yield a well-defined game with the desired influence, the approach adopted here has a significant advantage in simplicity.

The demand game just specified is the simplest generalization of Nash's perfect-information demand game with enough flexibility to make it *feasible* to

design incentive-efficient mechanisms for general environments. Although more realistic specifications of the rules of bargaining could easily be imagined, this specification is convenient, for several reasons. Like HM's unanimous-voting specification, it is simple and frictionless, and yields the "correct" conclusion about the limits of contracting when information is perfect or symmetric, or when information is asymmetric but agents can commit themselves to a mechanism before observing their private information.

Further, the generalized demand game provides a complete model of mechanism design, with all mechanism proposals endogenously determined, while allowing the widest possible range of welfare distributions to emerge in equilibrium. This kind of ambiguity is a natural feature for a determination of the limits of contracting to have, because the goal is to learn what kinds of outcomes the rules of contracting and the environment make it possible for rational agents to achieve. The factors that sharper predictions in particular cases must surely depend on—reputations, expectations in general, and bargaining skill, for instance—are excluded from the analysis by definition.

In HM's unanimous-voting analysis, on the other hand, the full range of welfare distributions is achieved only by varying the exogenous specification of the status quo mechanism. If one tried to adapt their framework to make all mechanism proposals endogenous, voting on proposals would have to be sequential, and the agent who made the first proposal would generally have some monopoly power, limiting the range of outcomes.

Before discussing the attainability restrictions implied by my specification of the rules of the mechanism-design process, two problems should be noted. First, since the rules of bargaining over mechanisms play such a crucial role in determining the attainability restrictions, it is disturbing that the demand game places arbitrary limits on agents' bargaining strategies. In particular, it assumes that agents are committed to terminate the search for a mechanism at a particular point, which trivializes issue 2. (Whether mechanisms themselves allow commitment is a separate issue, whose resolution is implicit in the set of feasible mechanisms.) It may be possible to base a more realistic determination of attainability restrictions on a more realistic specification of the rules of bargaining, e.g., along the lines suggested by Cramton (**2**, Chapter 5). This may be intractable for bargaining problems as complex as mechanism design, however.

Second, the rules that map agents' actions into physical outcomes in the demand game, even with perfect information, are highly sensitive to the details of the environment, and cannot be stated simply without reference to them. This complexity appears necessary, in multi-issue bargaining, to allow agents unilaterally to effect Pareto-improving adjustments in the terms of the agreement within a single period. Even though the rules of the demand game rely only on common-knowledge information, it can be argued that this sensitivity makes them an unlikely candidate for a social institution. Whether good welfare performance can be achieved with simpler mechanism-design rules is an interesting question, on which Kalai and Samet [**8**] have recently made some progress for abstract bargaining problems with perfect information. They consider rules in which

agents make simultaneous proposals directly in physical terms, and identical proposals signal an agreement. These rules are independent of the details of the environment, whose complexity is confined to agents' preferences over proposals. Kalai and Samet show that their rules, while compatible with almost any equilibrium outcome when agents have only one opportunity to make proposals, may ensure that the outcome is efficient if agents are allowed enough chances to recontract.

I shall now discuss the attainability restrictions implied by my specification of the rules of mechanism design. The set of decision rules that are attainable when mechanism design takes place ex ante is a useful benchmark. Then there can be no information leakage, so that once the set of feasible mechanisms is redefined to reflect incentive-compatibility constraints, the situation is formally analogous to contracting with perfect or symmetric information. It follows, recalling that incentive-compatibility includes interim individual rationality under my assumptions, that agents can agree on any ex ante incentive-efficient mechanism in a perfect mechanism-design equilibrium. Because there can be no information leakage, this implies that any ex ante incentive-efficient decision rule is attainable when mechanism design takes place ex ante. Other perfect-equilibrium outcomes are also possible, but they can be ruled out as in the perfect-information demand game. Thus, the demand-game specification yields the "correct" conclusion about the limits of contracting at the ex ante stage, in the strongest form that consideration of the perfect- and symmetric-information results suggests is possible.

Two important differences arise in interim mechanism design, because then an agent's types can use different mechanism-proposal strategies. (Of course, only the observed type actually makes a proposal when the game is played; the strategies of other types should be viewed as formalizations of other agents' expectations.) Although an agent must demand some utility level for each of his types, at the interim stage he cares only about how his observed type fares; this reflects issue 1. Further, if this induces an agent's types to use different mechanism-proposal strategies, other agents will draw inferences from his proposal; this is the essence of issue 3.

I shall now argue that these facts generally render some or all interim incentive-efficient decision rules unattainable when mechanism design takes place at the interim stage. It is helpful to distinguish between two kinds of mechanism-design equilibria. A *nonrevealing* equilibrium is one where each type of any given agent makes (or is expected to make) the same type-contingent demand. A *revealing* equilibrium can be defined, for my purposes, as any other equilibrium. With pure or mixed strategies, information leakage can occur only in revealing equilibria.

The intuitive reason why interim incentive-efficient decision rules are not generally attainable at the interim stage, for this section's specification, is as follows. Although the rules of mechanism design impose a mechanism without regard to whether it was generated in revealing or nonrevealing equilibrium, only nonrevealing equilibria can be expected to yield incentive-efficient decision rules in general. The demand game gives each agent the power to make changes that would be Pareto-improvements in the absence of information leakage. Only by

coincidence will this yield efficient outcomes when information leakage alters the incentive properties of the mechanisms that result from agents' demands. However, nonrevealing perfect equilibria are unlikely to exist here. The problem is that agents have too much freedom to "slant" their type-contingent demands in favor of their observed types at the expense of their other types. Because this can always be done without destroying compatibility, a type can unilaterally cause the selection of a large number of mechanisms that are potentially better for him than the one that would result from matching the demands made by his agent's other types. Some of these would necessarily be better if other agents drew no inferences from the defection—a zero-probability event, in the hypothesized non-revealing equilibrium—and even if other agents draw reasonable inferences, it is highly unlikely that no possible defections are beneficial for any agent. Revealing equilibria will therefore be the rule, with nontrivial attainability restrictions.

The resulting failure of interim incentive-efficiency is not surprising per se, given the analogous results in "ordinary" noncooperative models of bargaining with asymmetric information. A particular set of bargaining rules is unlikely to be an incentive-efficient mechanism for any reasonably general class of environments. But this failure occurs for somewhat novel reasons. Here, the set of incentive-efficient mechanisms is common knowledge throughout. The inefficiency arises in part because demands are not always compatible in revealing equilibria, so the mechanism-design process sometimes ends with no agreement, as in the inefficient, mixed-strategy equilibria in the perfect-information demand game. The other source of inefficiency is the effect of information leakage in the bargaining process on the performance of mechanisms; in ordinary bargaining, this would correspond to the process of bargaining over outcomes having an adverse effect on their quality.

I shall conclude this section with a discussion of the attainability of the equal-weights ex ante incentive-efficient decision rule in HM's [6, Section 6] example. To apply the demand game to HM's example, it is necessary to supplement their specification of the set of feasible agreements with an assumption about the consequences of playing the underlying game noncooperatively when no mechanism-design agreement is reached. For my purposes, any such specification that is worse than any feasible agreement for both types of each agent, no matter what inferences they draw, will do.

In HM's example, the unique equal-weights ex ante incentive-efficient direct-revelation mechanism specifies decision B whenever agent 2 reports type b and, otherwise, A if agent 1 reports type a and C if 1 reports b. With no information leakage, this mechanism yields expected utilities 3/2 for type 1a, 13/2 for 1b, 1 for 2a, and 1 for 2b in the truthful equilibrium; the associated decision rule is therefore attainable when mechanism design takes place at the ex ante stage. At the interim stage, it is natural to try to realize this decision rule by supporting the associated mechanism as a nonrevealing mechanism-design equilibrium. For this to be possible, there must exist specifications of the mechanisms that will be implemented when agents make compatible demands, and of agents' inferences

on observing events that have zero probability in the hypothesized equilibrium (all other inferences are determined mechanically from Bayes' Rule), that make all possible defections unprofitable.

It is clear that no type of either agent can benefit from a defection that renders agents' demands incompatible. There are, however, many potential opportunities for gain by defecting to compatible demands. Suppose, for instance, that if agent 1 demands (3/2, 13/2) and agent 2 demands (37/36, −67/36), the rules impose the mechanism that selects A and B each with probability 1/2 whenever 1 reports a, and A with probability 5/18 and C with probability 13/18 whenever 1 reports b. It is easy to verify that this mechanism is incentive-compatible and yields exactly the demanded type-contingent utilities at its truthful equilibrium. Because the outcome it imposes is independent of agent 2's report, its implications with information leakage can be evaluated without reference to agent 1's inferences about 2's type. Thus, if type $2a$ defects from the hypothesized nonrevealing equilibrium to the demands (37/36, −67/36), he can be sure of raising his expected utility from 1 to 37/36. It follows that the equal-weights ex ante incentive-efficient mechanism cannot be agreed upon in nonrevealing perfect equilibrium for this specification of the mechanism associated with the demands (3/2, 13/2) and (37/36, −67/36).

Although there seems little doubt that, as this example suggests, ex ante incentive-efficient decision rules are not generally attainable for this section's mechanism-design rules, the example is formally inconclusive for three reasons. First, there are usually many ways to choose a direct-revelation mechanism to implement a given configuration of demands, and my example shows only that one nonpathological way does not work. One would like to present an example in which no way works for at least some incentive-efficient decision rules. However, the performance of mechanisms in this context depends on agents' zero-probability inferences, and a generally agreed-upon standard of plausibility for such inferences has not yet been established. Nonrevealing equilibria plainly never exist, for this specification of the mechanism-design rules, when agents do not draw inferences from defections; but this seems implausible. What is plausible, and whether it is consistent with any nonrevealing equilibria, remain in doubt.

Further, because the existence of nonrevealing equilibria depends on the patterns of equilibrium outcomes for feasible mechanisms as agents' beliefs vary, one cannot make the usual argument that there is no loss of generality in using only incentive-compatible direct-revelation mechanisms to implement compatible demands. Indirect mechanisms generally allow a larger set of outcome patterns, and these might make it possible to support more nonrevealing equilibria. To put it another way, although there is a straightforward application of the Revelation Principle to the entire two-stage mechanism-design game, as in Section 2's argument that attainability implies incentive-compatibility, information-leakage problems make it impossible to use the Revelation Principle to restrict the set of mechanisms that might be selected within the game.

Finally, even if a given decision rule were known to be inconsistent with nonrevealing equilibrium, there remains the possibility that it arises coincidentally

in a revealing equilibrium, where the "wrong" mechanism is selected, but information leakage alters its properties to yield the desired result. This is unlikely, but computationally very difficult to rule out, even in simple examples.

4. A MODIFIED MECHANISM-DESIGN PROCESS

This section studies a simple modification of the demand-game specification of the rules that govern mechanism design studied in Section 3. The modification is designed to overcome information-leakage problems by allowing agents endogenously to restrict the language they can use in bargaining over mechanisms, leaving it rich enough to allow the design of efficient mechanisms, but coarse enough so that agents cannot benefit by making proposals that leak their private information.

Because agents' von Neumann-Morgenstern utility functions are type-contingent, there is no loss of generality in normalizing them so that the anticipated consequences of no agreement, with no information leakage, yield expected utility zero for each type of each agent. The modified mechanism-design process is the same as the one studied in Section 3, except that agents announce *weights* as well as demands. Agent i's weights, denoted $a^i = (a^i_1, \ldots, a^i_{T_{i-1}})$, are used to constrain the demands of agent $i-1$ (with agent 1's weights constraining agent n's demands) as follows. A vector is *proportional* to another vector if and only if one equals the other times a non-zero scalar. If agents' demands are compatible, and agent i's weights are proportional to agent $i-1$'s demands, $i = 1, \ldots, n$, then the associated mechanism is implemented as in Section 3, whether or not agents draw inferences during the mechanism-design process. (The factors of proportionality may differ across agents.) Otherwise the mechanism-design process ends with no agreement, and the underlying game is played noncooperatively, again without regard to agents' inferences. The definition of a nonrevealing equilibrium is extended to require identical weights as well as demands for all types of a given agent.

It is clear that this specification shares the advantages of Section 3's, with the exception that it is slightly more complex. In particular, it is a true generalization of the perfect-information demand game, and it makes any ex ante incentive-efficient decision rule attainable at the ex ante stage. Unlike Section 3's specification, it also has perfect equilibria at the ex ante stage that yield other interim incentive-efficient decision rules, even though this is undesirable; this could easily be fixed. I shall now prove two results about how it performs at the interim stage.

THEOREM 1. *In the modified mechanism-design process, any interim incentive-efficient decision rule is attainable when mechanism design takes place at the interim stage.*

PROOF: Any given interim incentive-efficient decision rule can be supported in nonrevealing perfect equilibrium as follows. Let each type of agent i, $i =$

1, . . . , n, demand the type-contingent expected utilities associated with the desired decision rule, and announce weights proportional to agent $i-1$'s demands. Let agents interpret any defection from this configuration of demands and weights as conveying no new information. Then any unilateral change in weights, with or without an accompanying change in demands, either has no effect on the weights' proportions and therefore no effect on the outcome, or causes the mechanism-design process to end with no agreement. No agreement results in noncooperative play of the underlying game with, by hypothesis, no new information. Because interim incentive-efficiency implies incentive-compatibility, hence interim individual rationality under my assumptions, this cannot be beneficial for the defecting type. A unilateral change in demands has the same effect, unless it leaves the demands proportional to the weights, controlled by another agent, that constrain it. An increase in demands, maintaining proportionality, destroys compatibility because the original configuration of demands was associated with an interim incentive-efficient decision rule. A decrease in demands, maintaining proportionality, conveys no new information to other agents, and therefore yields the defecting type lower expected utility than in the original configuration. It follows that no defections of any kind are beneficial to any type. Q.E.D.

Theorem 1 is the counterpart, for attainability defined in terms of the modified mechanism-design process, of HM's Theorem 4. The modified process overcomes the problems encountered in Section 3's specification by giving another agent control over the intrapersonal equity considerations that arise in determining a given agent's best demand. An agent is free to adjust his proposal in response to potential welfare gains, but cannot slant it in favor of his observed type without destroying the opportunity to reach agreement on a mechanism. It would also be possible to allow the weights to be specified exogenously or determined by an impartial referee, or to allow constraints more general than linear proportionality, without altering the conclusion of Theorem 1.

Theorem 1 leaves open the question of how many other decision rules are attainable at the interim stage for the modified mechanism-design process. Theorem 2 provides a partial converse to Theorem 1, showing that the modified process shares the perfect-information demand game's tendency to yield efficient outcomes, so that Theorem 1 is not vacuous.

Theorem 2 requires some new definitions. Let a *weakly* interim incentive-efficient mechanism (decision rule) be one for which there exists no incentive-compatible mechanism (decision rule) that simultaneously makes each type of each agent *strictly* better off. All interim incentive-efficient mechanisms are clearly weakly interim incentive-efficient, but not vice versa, in general. The two efficiency concepts are equivalent whenever it is always feasible to transfer small amounts of welfare across types and agents. Finally, the set of incentive-compatible decision rules will be said to be *comprehensive* if any interim individually rational decision rule that is not preferred, by any type of any agent, to a given incentive-compatible decision rule, is also incentive-compatible. Comprehensiveness is a standard assumption, but it is somewhat stronger than usual with asymmetric

information, because when it does not occur naturally, it requires the existence of disposal activities that may affect the incentive constraints.

THEOREM 2: *In the modified mechanism-design process, if the set of incentive-compatible decision rules is comprehensive, then any perfect, nonrevealing, pure-strategy, strongly interim individually rational equilibrium that can be supported with agents drawing no inferences from zero-probability events yields a weakly interim incentive-efficient decision rule.*

PROOF: The idea of the proof is a simple extension of the argument establishing the efficiency of pure-strategy perfect equilibria in the perfect-information demand game: the modified mechanism-design process allows each agent to adjust his demands, respecting the constraints imposed by his neighbor's weights, to realize any welfare gains that are consistent with his opponents' demands and incentive-compatibility. The need for the additional restrictions is illustrated in the proof, and will be discussed further below.

Suppose, by way of establishing a contradiction, that there exists an incentive-compatible decision rule strictly better for each type of each agent than the equilibrium decision rule. It is clear from the hypotheses of the theorem that no type can be using all zero weights or making demands that preclude compatibility and strong interim individual rationality. It therefore follows from comprehensiveness and the existence of such a decision rule that there also exists an incentive-compatible decision rule that is at least as good as the hypothesized equilibrium for all types of all agents and strictly better for some, and whose associated demands are in the same proportions as the hypothesized equilibrium demands. Thus, any type whose preference is strict can replace his original demands by his agent's type-contingent utilities at this decision rule, without changing his weights. This preserves compatibility and, since other agents draw no inferences from the defection by hypothesis, benefits the defector, contradicting the hypothesis that the original configuration is an equilibrium. Q.E.D.

The hypotheses of Theorem 2 are strong, but the result nevertheless suggests that the modified mechanism-design process aggregates agents' preferences over mechanisms at the interim stage almost as effectively as the simpler process studied in Section 3 does at the ex ante stage, or as the ordinary demand game does with perfect or symmetric information. I shall now comment on the various restrictions. They all involve agents' expectations, which cannot be controlled by mechanical modifications of the mechanism-design process that preserve its flexibility. Perfectness guarantees equilibrium play once the mechanism is selected and that agents draw the correct inferences from first-stage proposals. Strong interim individual rationality rules out trivial (but not necessarily imperfect) equilibria where some or all types announce zero weights; these might also be eliminated by refining the perfectness requirement, but the present approach is simpler.

The restriction to pure-strategy equilibria serves precisely the same purpose as in the perfect-information demand game. Here, because the inefficient lack of coordination of demands associated with mixed strategies can be duplicated, with asymmetric information, when different types of a given agent use different strategies, it is also necessary to rule out revealing equilibria. The reader who is familiar with Myerson's [11] analysis of the informed-principal problem will already have noted that nonrevealing strategies are analogous to his concept of *inscrutable* mechanism proposals by the principal. (Inscrutability requires that all types of the principal propose the same type-contingent mechanism.) In [11], an argument that restricting consideration to inscrutable proposals involves no loss of generality is based on two observations. First, because only the principal's observed type ever gets to make a proposal, other agents can never refute the hypothesis of inscrutability by observation. Second, because the principal moves first and other agents' inferences depend only on observables, he can achieve with an inscrutable proposal any outcome that is possible with a scrutable proposal, by communicating the information that such a proposal would leak within the proposed mechanism. The first of these observations remains valid in specifications based on the demand game, but the second does not. Here, different sets of self-confirming expectations are possible with revealing and nonrevealing strategies, so the restriction to nonrevealing equilibria may involve some loss of generality.

In defense of the restriction to nonrevealing, pure-strategy equilibria, it can be said that uncertainty about the demands and weights agents intend to announce, while it can be consistent with rational expectations, serves no strategic purpose here. A suggestion from a referee that agents should employ pure, nonrevealing strategies would have a powerful self-enforcing quality, eliminating avoidable inefficiency without introducing bias into the outcome. Further, because the hypothesis that an agent is playing a nonrevealing strategy cannot be refuted by observation, it is best viewed as a convenient way to describe other agents' inferences from his proposals. When agents' inferences are described in a nonrevealing equilibrium, their plausibility must be determined by examining the zero-probability-updating rules that describe how defections from that equilibrium are interpreted.

Suppose that an agent expects that other agents will draw no inferences from defections, and that the weights that constrain his demands will be generated by a nonrevealing weight strategy, in which all types of his neighbor announce the same weights. Then it is easy to verify that the linear proportionality of demands and weights required for agreement and the von Neumann-Morgenstern structure of agents' risk preferences imply that all types of that agent have identical preferences over all possible strategies, even in the face of uncertainty about the compatibility of their demands with other agents' possibly uncertain demands. Thus, if plausibility is taken to require that zero-probability-updating rules interpret a defection as evidence, if anything, in favor of the type or types for whom the defection is least costly relative to the equilibrium (see, for example, Kreps and Wilson [9, Section 8]), then drawing no inferences from defections from a

nonrevealing, pure-strategy equilibrium is plausible, possibly uniquely so. Plausibility might be consistent with other inferences in revealing equilibria, or even in nonrevealing equilibria, but this argument suggests that the restriction on zero-probability-updating rules assumed in Theorem 2 is reasonable, given the other hypotheses.

5. CONCLUSION

This paper makes three main points. First, a full understanding of the limits of contracting at the interim stage must rest on a complete specification of the rules that govern mechanism design, because those rules influence how agents interpret other agents' mechanism proposals, and those interpretations in turn affect the incentives created by the agreed-upon mechanism. Second, for a game-theoretic analysis of endogenous mechanism design to be possible, the rules of bargaining over mechanisms must be specified by the modeler rather than derived within the model. Finally, some specifications of those rules imply limits of contracting at the interim stage analogous to, but generally different from HM's [6] durability restrictions; but at least one simple specification eliminates such restrictions, making it possible for agents to achieve, at the interim stage, the same incentive-efficient results that are attainable in frictionless bargaining over mechanisms at the ex ante stage. Because no specification can eliminate incentive-compatibility constraints, this is the best welfare performance possible at the interim stage, and is weakly interim Pareto-superior to that of any other mechanism-design process. This specification works by allowing agents to deal with information-leakage problems by endogenously restricting the language of bargaining in a way that leaves it rich enough to allow efficient aggregation of their preferences over mechanisms, but coarse enough to prevent them from encoding their private information in mechanism proposals.

Two directions for future work seem especially promising. First, the limits of contracting might be evaluated with more satisfactory models of noncooperatively rational, frictionless bargaining, perhaps along the lines suggested by Cramton [2] or Kalai and Samet [8]. As I have argued above, the fact that the set of incentive-efficient mechanisms is common knowledge suggests that the mechanical rules of bargaining in such models should not be the sole determinant of how agents share the gains from contracting. But this line of research will ultimately require closer attention to other kinds of multiple-equilibrium problems.

The second direction concerns the standards for determining whether a mechanism-design process is workable. I have adopted the standard assumption that any process that can be specified using only information that is common knowledge to agents is usable. However, in general environments, achieving the best possible welfare performance in mechanism design at the interim stage requires bargaining rules that are unrealistically sensitive to the details of the environment. In fact, both specifications of the mechanism-design process studied here seem likely to require the assistance of a referee, and are unlikely to serve as social institutions

whose rules can be transmitted and implemented by agents unaided. It would be of great interest to know what sorts of environments imply that there are little or no welfare costs of requiring mechanism-design rules, or mechanisms themselves, to be independent of the details of the environment; Wilson [13] has obtained some results along these lines for a double actions in simple environments. Perhaps, as I have argued in [4], simple social institutions tend to emerge when these costs are low, and more complex institutions like arbitration are used when the costs are high enough to justify designing a new mechanism for each environment.

University of California, San Diego

Manuscript received September, 1983; final revision received August, 1984.

REFERENCES

[1] BINMORE, K.: "Nash Bargaining and Incomplete Information," Economic Theory Discussion Paper 45, Cambridge University, 1981.
[2] CRAMTON, P.: "The Role of Time and Information in Bargaining," unpublished Ph. D. dissertation, Graduate School of Business, Stanford University, June, 1984.
[3] CRAWFORD, V. P.: "A Theory of Disagreement in Bargaining," *Econometrica*, 50(1982), 607–637.
[4] ———: "The Role of Arbitration and the Theory of Incentives," to appear in *Game-Theoretic Models of Bargaining*, ed. by A. Roth. Cambridge: Cambridge University Press, 1985.
[5] HARSANYI, J.: "Analysis of a Family of Two-Person Bargaining Games with Incomplete Information," *International Journal of Game Theory*, 9(1979), 65–89.
[6] HOLMSTROM, B., AND R. MYERSON: "Efficient and Durable Decision Rules with Incomplete Information," *Econometrica*, 51(1983), 1799–1819.
[7] KALAI, E., AND R. ROSENTHAL: "Arbitration of Two-Party Disputes under Ignorance," *International Journal of Game Theory*, 7(1978), 65–72.
[8] KALAI, E., AND D. SAMET: "Unanimity Games and Pareto Optimality," Northwestern University, Center for Mathematical Studies in Economics and Management Science, Discussion Paper 546, January, 1983.
[9] KREPS, D., AND R. WILSON: "Sequential Equilibria," *Econometrica*, 50(1982), 863–894.
[10] MYERSON, R.: "Incentive Compatibility and the Bargaining Problem," *Econometrica*, 47(1979), 61–73.
[11] ———: "Mechanism Design by an Informed Principal," *Econometrica*, 51(1983), 1767–1797.
[12] NASH, J.: "Two-Person Cooperative Games," *Econometrica*, 21(1953), 128–140.
[13] WILSON, R.: "Efficient Trading," Stanford University, IMSSS Technical Report 432, October, 1983.