INSTITUTE FOR ECONOMIC RESEARCH

Discussion Paper # 952

Extensive Form Implementation in Incomplete Information Environments

by

James Bergin and Arunava Sen

September 1997

Queen's University
Kingston, Ontario, Canada K7L 3N6

Extensive Form Implementation in Incomplete Information Environments*

James Bergin
Department of Economics
Queen's University
Kingston, ON K7L 3N6
Canada

Arunava Sen
Indian Statistical Institute
S.J.S Sansanwal Marg
New Delhi 110016
India

January 1997

* We thank John Duggan and Matt Jackson for comments on earlier drafts. We also thank an anonymous referee for many helpful comments. Bergin acknowledges financial support from the Social Sciences and Humanities Research Council of Canada and the Advisory Research Council at Queen's University. All errors are our own.

Abstract

We study the implementation of social choice rules in incomplete information environments. A sufficiency condition called *posterior reversal* is given for extensive form implementation. The condition has a natural interpretation in signaling terms: consistent posterior distributions under truth-telling are different from consistent posteriors under deception. This variation in the distribution over player types leads to variation in the distribution over actions and outcomes (comparing truth-telling and deception). We exploit this feature to implement social choice rules.

1 Introduction.

When agents interact within an institutional framework, the structure of the institution plays a central role in determining outcomes: it determines the choices agents can make, the strategic considerations involved, and how actions translate into outcomes. For example, in the context of voting, when a group of individuals must select some alternative, voting by veto, majority rule, and so on, are alternative institutions that select outcomes. The design of institutions which produce desirable outcomes (by some criteria) is a fundamental problem of economic theory.

The theory of implementation and mechanism design is concerned with institutional design in situations where outcomes in an environment are exogenously given and where the objective is to specify rules of interaction between agents that lead to desirable outcomes. What constitutes a desirable outcome will typically depend on the characteristics of the agents involved (the state), varying as these characteristics vary. For example, in designing a voting procedure to select between two candidates, a common requirement is that the chosen candidate be top ranked by at least half the voters. Here, at any state the desired outcome (among two alternatives) is that preferred by the majority. In this case, as preferences of voters change, so does the best choice of candidate. If the institution is to survive over time and produce appropriate outcomes as participants (preferences) vary, the institutional details should be independent of the characteristics of individuals. The institution must, for a given collection of participants, generate strategic considerations that lead the interaction of those agents to a desired outcome, and as preferences vary, the institution must generate new and appropriate strategic considerations, leading to the correct outcome at the new state. Given a rule associating outcomes to states (a social choice rule), this implicitly imposes requirements on the variation of preferences across states in order to structure incentives appropriately, so that as preferences vary the outcome generated by the institution varies accordingly. This leads to the central difficulty arising in the design of mechanisms in implementation theory, and when a mechanism exists with equilibrium outcomes that track the social choice rule as states vary, the choice rule is said to be implementable.

Different forms of mechanism reflect different institutional features. Thus, for example, majority voting is a situation in which agents move simultaneously and the outcome is then selected. This is naturally modeled as a strategic form game. Voting by veto highlights important temporal features – the choice made by one agent may restrict the choices of others, and is most naturally modeled as multi-stage game.

For mechanisms modeled as strategic form games the central characterizing property of implementable social choice rules is monotonicity. Monotonicity is a necessary property of a social choice rule for the rule to be implementable in a strategic form mechanism. Monotonicity was introduced by Maskin (1977) and requires that if an outcome is selected by the social choice rule at some state and in moving to another state the outcome falls in no agent's ranking, then the social choice rule should select the same outcome at the new state. The intuition is natural: if the outcome doesn't fall in anyone's ranking and was previously chosen, then since no one previously wished to challenge the outcome they will not do so in the new state where it is now ranked at least as highly. Alternatively, if the outcome selected by the social choice rule varies in moving from one state to another it must be that at the new state the outcome chosen at the old state is ranked lower than it was at the old state (relative to some alternative), for some individual. Phrased this way, the requirement highlights the need for sufficient variability in preferences relative to the social choice rule. Maskin's monotonicity condition applies to complete information environments. In the context of incomplete information environments, which we focus on here, there is an analogous condition called Bayesian monotonicity. Bayesian monotonicity was identified by Postlewaite and Schmeidler (1986) as a necessary

condition for strategic form implementation. Jackson (1991) provides necessary and sufficient conditions for strategic form implementation in incomplete information economic environments. The problem of finding necessary and sufficient conditions in general incomplete information environments is far more difficult. See, for example, Dutta and Sen (1994 and 1995). Subsequent to the work on strategic form implementation, research proceeded in at least three directions. Palfrey and Srivastava (1989) consider strategic form games, and focus on a refinement of Nash equilibrium, excluding those equilibria in which some agent plays a weakly dominated strategy. Abreu and Sen (1991), Abreu and Matsushima (1990) and Matsushima (1990) reformulate the problem by working with lotteries on outcomes and exploit the linearity of vonNeumann-Morgenstern preferences. Extensive form implementation in subgame perfect equilibrium is considered by Abreu and Sen (1991) and Moore and Repullo (1988).

Use of extensive forms is particularly well suited to incomplete information implementation since these games provide a natural framework for signaling and information transmission. Recently, Baliga (1993), Bergin and Sen (1993), and Brusco (1995) have considered the impact of using extensive form games to implement social choice rules. In the complete information context, subgame perfection is a natural choice for the solution concept. In incomplete information environments, the analogous requirement is sequential rationality — that players should make optimal choices whenever called on to move. Sequential rationality is standard in most solution concepts[1] and requires that agents make rational choices relative to some belief system. The specification of beliefs is implicitly determined by the choice of solution concept (such as sequential equilibrium, perfect Bayesian equilibrium, equilibrium based on forward induction belief restrictions, and so on). However, because equilibria are sensitive to belief specification, whether a game implements a given social choice rule or not depends critically on how belief restrictions are imposed. We approach this problem by focusing on beliefs (posterior distributions) that are sufficient to permit implementation, *independent* of the solution concept. Apart from having the advantage that the conditions are independent of the equilibrium concept (given sequential rationality), the approach allows the results to be interpreted in terms of standard signaling ideas. We primarily consider games that have no equilibria that go beyond a first stage. (This is conventional in complete information extensive form mechanisms.) We refer to such a game as a game with one round of signaling. In the context of incomplete information a significant virtue is that it avoids the difficulty of having to track sequences of posterior distributions and inevitable dependence of the implementation on the precise manner in which a solution concept restricts beliefs.

A key insight of the recent literature on implementation theory is the crucial role played by appropriate preference reversals in successful implementation. In normal form implementation this reversality requirement is called Bayesian monotonicity and postulates the existence of an allocation that undergoes preference reversals vis-à-vis the allocations that arise under truth-telling and deception. An important feature of these reversals is that they occur with respect to the prior distribution on types. Allowing for multiple stages in the game form permits a significant weakening of the Bayesian monotonicity requirement: we can now additionally use posterior distributions to generate reversals, and we can create more finely tuned incentives in the multistage framework. For example, a type report when an agent reports "truthfully" conveys different information than the same report when the agent reports "dishonestly", and hence may lead to different behavior.

To motivate the signaling role of beliefs in the simplest possible way, consider a situation where a player

[1] Baliga uses perfect sequential equilibrium, Brusco adopts perfect Bayesian equilibrium and Bergin and Sen (1993) use sequential equilibrium.

has two possible and equally likely types, a and b. Suppose that the player's strategy is to announce a type. Suppose also that one strategy (say $\hat{\alpha}$) announces a when the player is type a and announces b when the player is type b. Another strategy, $\tilde{\alpha}$, announces b when the player is type a and announces a when type b. Under $\hat{\alpha}$, the posterior distribution over types given the announcement b is that the agent is type b with probability 1, while under $\tilde{\alpha}$ and given b, the posterior distribution puts probability 1 on type a. Let $\mathbf{P}_\delta(\cdot \mid y)$ denote the posterior distribution conditional on y when strategy δ is used, so that $\mathbf{P}_{\hat{\alpha}}(a \mid b) = 0$ and $\mathbf{P}_{\tilde{\alpha}}(a \mid b) = 1$. Thus, the posterior distributions are distinct.[2] In this discussion Bayesian updating is used — the relevant event in each case has positive probability. Generally, in multistage games in deriving beliefs two complications arise: it's necessary to consider the structure of beliefs on zero probability events and secondly, with many agents and correlated types, these beliefs are not uniquely defined when conditioning on events that have zero probability. Nevertheless, it turns out that properties similar to those described above are preserved in belief systems. We exploit this fact in a game form to generate different behavior in subforms of the extensive form, depending on the strategies used by agents (in particular "truth-telling" and "deception" strategies). From a technical perspective, there are two ways in which this variation in behavior may occur. A change in the distribution over nodes in a subform may change the distribution over outcomes simply because different choices are made at different nodes; and a change in the distribution over nodes of the subform may alter the choices made at a given node.

In section 2 we describe the model and introduce extensive form implementation in section 3. There, we provide three examples that are central in motivating the discussion. The first two illustrate the ideas described above (on the impact of variation in distributions over subform nodes). The third example gives a social choice function that can be implemented in the extensive form but fails the Abreu-Matsushima necessary condition (measurability) for implementation in the normal form. This is in contrast to the complete information case where normal form refinements are more powerful than subgame perfection in the extensive form. The example is constructed so that the interim expected utilities are *type independent* (although the ex post utilities depend on the type profiles). As a result, the only Abreu-Matsushima measurable functions are constant on the type space, and any social choice function that varies over types fails measurability. Implicitly, the measurability of a social choice function depends on the *prior* distribution over types via the interim expected utilities. Implementation is possible in the extensive form by shifting the problem to a subform where posterior distributions are such that different types of an agent have different incentives as the (consistent) posterior distributions vary between truth-telling and deception. In section 4 we describe the posterior reversal condition and give a sufficiency condition for extensive form implementation. A central requirement of posterior reversal is that for any deception there is some signal such that the posterior under truth-telling at that signal is distinct from the posterior under deception (in fact at out of equilibrium signals we deal with sets of distributions, so the condition is formulated in terms of disjointedness of sets of posterior distributions.) We illustrate the condition with the examples and give some simplifications of the condition for special cases. In addition, we contrast the ideas in posterior reversal with Bayesian monotonicity and some extensions. In section 5 we conclude with a discussion of the literature.

2 The Model.

The set of agents is denoted I, with $I = \{1, 2, \ldots, n\}$. In the incomplete information environment, each

[2] Note that this observation remains valid if mixed strategies are used. Suppose that an arbitrary strategy, γ, puts positive probability on b when the agent is type a. Then the posterior distribution under gamma puts positive probability on a; given the announcement b, $\mathbf{P}_\gamma(a \mid b) > 0$, whereas under the truth-telling strategy $\hat{\alpha}$, $\mathbf{P}_{\hat{\alpha}}(a \mid b) = 0$.

agent i has a set of types, S_i. In addition, a fixed prior distribution, μ, over types $S = \times_{i=1}^{n} S_i$ is given. We assume that S is finite and that for each $s \in S$, $\mu(s) > 0$. The set of outcomes is denoted A. An allocation is a rule that assigns outcomes to types. Formally,

Definition 1 *An* **allocation** *is a function* $x : S \to A$.

Denote by X the set of all allocations. The utility function of player i is a function from $A \times S$ to $\mathcal{R}$, $u_i : A \times S \to \mathcal{R}$. Given an allocation x and a distribution μ on S (the prior distribution), the expected utility of agent i conditional on type $s_i \in S_i$ is $V_i(x, s_i \mid \mu) = \sum_{s_{-i} \in S_{-i}} u_i(x(s), s)\mu(s_{-i} \mid s_i)$. (Given a vector x, x_{-i} is obtained from x by deleting the ith component.) A binary relation on allocations (X) is then defined according to: $x\mathbf{R}^i(s_i, \mu)y \Leftrightarrow V_i(x, s_i \mid \mu) \geq V_i(y, s_i \mid \mu)$. When the inequality is strict write $x\mathbf{P}^i(s_i, \mu)y$.

Definition 2 *A* **social choice correspondence** *(SCC), F, is a subset of X, the set of allocations. In the case where the SCC contains just one element of X so $F = \{x\}$, it is called a* **Social Choice Function** *(SCF).*

Intuitively, the outcome $x(s)$ is the preferred social outcome at player type profile s. However, the types vector s is not observable. In a mechanism where agents report their types with report $\hat{s}$ leading to outcome $x(\hat{s})$, then under truthful reporting the desired outcome is achieved. In general, agents incentives may conflict with truthful reporting and this in turn may lead to deceptions by agents, reporting types other than the true type.

Definition 3 *A* **deception for** i *is a* $\alpha_i : S_i \to S_i$, $\alpha_i \in D_i = \{\tilde{\alpha}_i : S_i \to S_i\}$. *A* **deception** *is a function* $\alpha = \{\alpha_i\}_{i=1}^{n}$, $\alpha \in D = \times_{i=1}^{n} D_i = \{\tilde{\alpha} \mid \tilde{\alpha} : S \to \times_{i=1}^{n} S_i\}$.

Denote by $\hat{\alpha}_i$ the identity function on S_i and by $\hat{\alpha}$ the identity function on S. Given an outcome function x and $\alpha \in D$, define x_α: $x_\alpha(s) \equiv x(\alpha(s))$, $\forall s \in S$. Thus, if agent i adopts the reporting strategy $\hat{\alpha}_i$, then the agent plans to report truthfully. If the agent adopts the reporting strategy $\alpha \neq \hat{\alpha}$, then for some type of s_i of i, the player plans to report dishonestly, reporting, say, $\tilde{s}_i = \alpha_i(s_i) \neq s_i$. When, for each i, truthful reporting by other agents leads to it being optimal for i to report truthfully, then the social choice function satisfies "self selection". Formally,

Definition 4 *The social choice function $F = \{x\}$ satisfies* **self selection** *if*

$$x\mathbf{R}^i(s_i, \mu)x_{(\tilde{\alpha}_i, \hat{\alpha}_{-i})}, \forall \tilde{\alpha}_i \in D_i, s_i \in S_i, i \in I.$$

Self selection (or incentive compatibility) is an essential requirement of "implementable" social choice functions. If the social choice function satisfies self-selection, then in a type announcement game (a "direct mechanism") truth-telling is a Nash equilibrium in the game where agents receive $x(s)$ when the announcement is s. If the social choice function fails self-selection then there is no mechanism (direct or otherwise), in which the equilibrium leads to the outcome s at type profile s, for all $s \in S$ (Myerson (1979)).

Remark 1 Throughout the paper we focus on pure strategies and restrict attention to pure strategy equilibria. This is conventional although not universal. The concept of Bayesian monotonicity (see below) is central in the literature and is defined for deceptions that do not involve randomization. Working with pure strategies has the virtue that we maintain comparability with much of the existing literature. However, many of our characterizations require no reference to the issue of pure versus mixed strategies. Since the distinction is important, we will indicate where the distinction is not significant or does not play a crucial role.

2.1 Normal Form Implementation.

Mechanism design in incomplete information environments has been studied by Postlewaite and Schmeidler (1986), by Palfrey and Srivastava (1989) and more recently by Jackson (1991). A key necessary condition for implementation in normal form games in Nash equilibrium is Bayesian monotonicity.

Definition 5 *A social choice function x satisfies* **Bayesian monotonicity** *if given $\alpha \in D$, $x_\alpha \neq x$, $\exists i \in I, s_i \in S_i$ and $y \in X$ such that:*

1. $xR^i(t_i, \mu)y_{\alpha_i(s_i)}$, $\forall t_i \in S_i$
2. $y_\alpha P^i(s_i, \mu)x_\alpha$

where $y_{\alpha_i(s_i)}(t) = y(t_{-i}, \alpha_i(s_i))$, $\forall t \in S$.

Remark 2 When Bayesian monotonicity holds, conditions 1 and 2 are satisfied at every α (with $x_\alpha \neq x$). When 1 and 2 hold at some given α, we say that Bayesian monotonicity is satisfied at α.
Motivation for this condition is given in Palfrey and Srivastava (1989) and Palfrey (1992). Necessary and sufficient conditions for Nash implementation in *economic environments* are given in Jackson (1991). Given $x, y \in X$, and $T \subseteq S$, define $x_T y = \chi_T x + (1 - \chi_T)y$, where χ_T is the indicator function of T. Call $\{\{S_i\}_{i=1}^n, \{u_i\}_{i=1}^n, A, X\}$ an *environment.*

Definition 6 *An environment is* **economic** *at μ if given $z \in X$ and $s \in S$, $\exists i, j \in I$ $(i \neq j)$, $x, y \in X$ x, y both constant with $(x_T z)P^i(s_i, \mu)z$ and $(y_T z)P^j(s_j, \mu)z$, $\forall T \subset S$ such that $s \in T$.*
Say that an environment is economic if for each $\mu \in \Delta(S)$, the environment is economic at μ.
In "economic environments" with $n \geq 3$, a social choice function is implementable in Nash equilibrium if and only if it satisfies both self selection and Bayesian monotonicity (Jackson (1991)).

3 Extensive Form Implementation

We denote an extensive form game of incomplete information by Γ. A detailed description of extensive form games is given in Selten (1975), and is too involved to fully review. Here we give a minimal description of the relevant components. A path in the game is a state (type) profile $s \in S$ and a history, h, of actions chosen by the players. Thus a path is a pair (s, h). The set of all paths is denoted $S \times H$. A payoff for each player is associated to each path: $\pi(h, s) = \{\pi_i(h, s)\}_i$, where $\pi_i(h, s)$ is the payoff to player i if the type profile is s and the action history is h. An information set for player i is a subset of the set of histories, identifying those paths indistinguishable to the agent when required to move and a type, $s_i \in S_i$. Let $\sigma_i \in \Sigma_i$ be a behavioral strategy for player i in Γ, specifying an action choice at each information set of i. Thus, a strategy for i can be written as $\{\sigma_i(s_i)\}_{s_i \in S_i}$, where $\sigma_i(s_i)$, specifies an action for i, type s_i at each information set. Given a strategy, $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \in \Sigma = \times_i \Sigma_i$, and a realization $s \in S$, a payoff to agent i is determined, say $\pi_i^*(\sigma, s)$. A prior distribution μ on S determines an expectation operator $\mathbf{E}_\mu$ such that the expected payoff to player i under strategy σ is $\mathbf{E}_\mu\{\pi_i^*(\sigma, s)\}$ and the conditional payoff to player i, type s_i is $\mathbf{E}_\mu\{\pi_i^*(\sigma, s) \mid s_i\}$. Alternatively, note that each strategy, σ, and type distribution μ determine a distribution, $\varphi_{(\sigma,\mu)}$, on $H \times S$ with associated expectation operator $\mathbf{E}_{(\sigma,\mu)}$. Formulated in this way, the expected payoff to player i is $\mathbf{E}_{(\sigma,\mu)}\{\pi_i(h, s)\}$, and the expected payoff to player i type s_i is $\mathbf{E}_{(\sigma,\mu)}\{\pi_i(h, s) \mid s_i\}$. Let $\mathcal{I}_i$ denote the collection of information sets of player i. If σ^* is a Nash equilibrium and the information set $I_i \in \mathcal{I}_i$ is reached with positive probability given the strategy σ^* and type distribution μ, then:

$$\mathbf{E}_{(\sigma^*,\mu)}\{\pi_i(h, s) \mid I_i\} \geq \mathbf{E}_{(\{\sigma^*_{-i}, \sigma_i\},\mu)}\{\pi_i(h, s) \mid I_i\}, \ \forall \ \sigma_i \in \Sigma_i.$$

This condition is called *sequential rationality* at I_i and any Nash equilibrium strategy of the strategic form of the game induces sequential rationality along the equilibrium path. Beliefs on the equilibrium path are determined by Bayes' rule. However, it is necessary to model rational behavior of a player faced with an unanticipated decision problem (i.e. choosing optimally off the equilibrium path). If I_i has probability 0 under (σ, μ), an "appropriate" conditional distribution is assigned. Different solution concepts (perfection, sequential equilibrium, and various forms of perfect Bayesian equilibrium) develop different procedures for restricting out of equilibrium beliefs. This raises the difficulty that equilibrium outcomes of a given extensive form may vary with the solution concept and is obviously important for the implementation problem. We will return to this issue later. For now, we fix some equilibrium criterion such as sequential equilibrium or perfect Bayesian equilibrium and use the term "equilibrium" to mean that some such solution concept has been adopted and the equilibrium is in terms of the solution criterion.

Definition 7 *An* **extensive form mechanism** *is an extensive form game of incomplete information, Γ, with type space $\times_{i=1}^{n} S_i$, prior distribution μ over types, and outcomes determine by histories according to a rule $a : H \to A$.*

That a depends only on H, and not $H \times S$, reflects the fact that the mechanism cannot depend on unobservables (the player types). Thus, to each path in the game the mechanism associates an outcome. If the player type vector drawn by μ is s, then the payoff to agent i is $u_i(a(h), s)$. Thus with $\pi_i(h, s) = u_i(a(h), s)$, the expected payoff to i is $\mathbf{E}_{(\sigma,\mu)}\{u_i(a(h), s)\}$, and the expected payoff to player i type s_i is $\mathbf{E}_{(\sigma,\mu)}\{u_i(a(h), s) \mid s_i\}$.

Given the extensive form game Γ, let Ψ_Γ be the set of equilibrium strategies, or write $\Psi_\Gamma(\mu)$ to make explicit the dependence on μ. Thus, $\Psi_\Gamma : \Delta(S) \to \Sigma$. Let $\sigma^* \in \Psi_\Gamma(\mu)$. A strategy profile and the prior distribution, (σ, μ), determine a distribution over histories, which we denote φ. For each $s \in S$ a conditional distribution on A is determined according to

$$\phi_{(\sigma^*,\mu)}(B \mid s) = \varphi_{(\sigma^*,\mu)}(\{h \in H \mid a(h) \in B\} \mid s), B \subseteq A.$$

Let $supp\ \phi_{(\sigma^*,\mu)}(B \mid s)$ be the support of this distribution in A.

Definition 8 *The allocation $x : S \to A$ is* **implementable in the extensive form** *in sequential equilibrium if there exists an extensive form mechanism Γ (with prior distribution μ on S), such that the game has a sequential equilibrium and such that if σ^* is a sequential equilibrium, then $x(s) = supp\ \phi_{(\sigma^*,\mu)}(B \mid s), \forall s \in S$.*

Remark 3 Implementation in perfect Bayesian equilibrium is defined by replacing "sequential equilibrium" with "perfect Bayesian equilibrium" in this definition. Implementation in other extensive form concepts is defined analogously. When we need to work with a specific solution concept, we use sequential equilibrium. Other solution concepts that impose sequential rationality at every information set would do equally well.

3.1 Examples.

In this section we discuss three examples. The first example shows how the majority rule social choice function can be implemented in a public goods problem where Bayesian monotonicity fails. We provide a complete description of the (simple) implementing game form. The key feature of this example is that in moving from truth-telling to deception, the posterior distribution over types of some agent varies. As long as different types act differently, this gives a different distribution over outcomes, comparing truth-telling to deception. The second example illustrates how the variation in the posterior (between truth-telling and

deception) can cause the same type of a player to play differently. This is again an example of posterior reversal, but also of chain reversal. The third example shows how complex strategic behavior generated by variation in beliefs may be exploited. We choose a social choice function that violates Abreu-Matsushima measurability and show how it can be implemented in the extensive form. The example is constructed so that interim expected utilities are *type independent* (making it impossible to elicit distinct type-dependent behavior). The example shows that it is sometimes necessary to use extensive form games to implement a social choice function.

3.1.1 Example 1: Implementing a Non-monotonic SCF

The following example is discussed in Palfrey and Srivastava (1989) and Palfrey (1992). There are three players and each player has two types, $S_i = D = \{a, b\}, i = 1, 2, 3$. Types are independently drawn: a with probability μ_a and b with probability $\mu_b = 1 - \mu_a$. Preferences are given by: $u_i(d, s_i) = 1,\ s_i = d \in \{a, b\}$ and $u_i(d, s_i) = 0,\ s_i \neq d \in \{a, b\}$. Whatever "state" $s \in S = \times_{i=1}^{3} S_i$ is realized, at least two agents are drawn with the same type. The social choice function is majority rule: $x(s_1, s_2, s_3) = d$, if $\exists i \neq j, s_i = s_j = d$. This allocation fails Bayesian monotonicity for some values of the prior distribution (μ_a, μ_b). Hence, at such prior distributions, the social choice function cannot be implemented as a Nash equilibrium in a normal form game. However, this social choice function can be implemented in an extensive form game as follows. Define an extensive form game with two stages. In the first stage each player announces their type. In addition, player 1 announces an element of $\{c, n\}$. If agent 1 announces n, then the game terminates at stage 1 with the majority announcement selected: if two agents choose $d \in \{a, b\}$, then d is selected. If agent 1 announces c and if either of the profiles (a, a, a) or (b, b, b) are announced, then the game goes to stage 2 where player 2 is allowed to choose the outcome in $\{a, b\}$.

First, truth-telling is an equilibrium. For each type of each player, truth-telling weakly dominates non truthful reporting. If agent one announces c, the game goes to stage 2 when either (a, a, a) or (b, b, b) are announced, but in the first case player 2 picks a (since given 2 announces a, 2 is type a with probability 1) and in the second b, so the outcome is unaffected. Thus truthful reporting by all players, player 1 selecting n in period 1 and 2 selecting his type if stage 2 is reached forms an equilibrium.

Next, there is no deception equilibrium where both a and b are announced with positive probability. To see this, suppose otherwise. Then $\exists j, k$ such that $p(a \mid j)p(b \mid k) > 0$ (when $j = k$ each type of j announces differently). In this case player $r \notin \{j, k\}$ faces a distribution where both a and b have positive probability of being announced by other players and so has a unique best response – reporting truthfully.

The only remaining possible "deception" equilibrium is one where $p(d \mid 1)p(d \mid 2)p(d \mid 3) = 1$ for some $d \in \{a, b\}$. Suppose that this holds for $d = a$. In this case the (consistent) posterior distribution over player 2's types coincides with the prior (since $\sigma_a^2(a) = \sigma_a^2(b) = 1$, where $\sigma_d^i(s_i)$ is the probability that player i type s_i announces d).

$$P(b \mid a) = \frac{\sigma_a^2(b)\mu_b}{\sigma_a^2(b)\mu_b + \sigma_a^2(a)\mu_a} = \frac{1\mu_b}{1\mu_b + 1\mu_a} = \mu_b$$

When player 1 plays $\sigma_a(a) = \sigma_a(b) = 1$ and each type of 1 chooses n, then neither player 2 or 3 has any incentive to defect. However, if agent 1 type b "defects" to the signal c then conditional on agent 1 being type b, the game goes to stage 2 with probability 1, since a is played with probability 1. Here, the posterior probability on agent 2's type is the prior, so that with probability μ_b, agent 2 will choose b. Thus, the defection leads to b being chosen with probability μ_b. Since $\mu_b > 0$, agent 1 type b has the incentive to defect, thus upsetting these strategies as equilibrium strategies. Finally, note that there cannot be an

equilibrium where in the first stage each agent announces (independent of type) (a, a, a), say, and agent 1 type b announces c. In this case the game goes to stage two with positive probability (μ_b), and conditional on reaching stage two, b is chosen with positive probability (again μ_b). This cannot be an equilibrium because agent 3 type a has an incentive to announce b: if agent 3 type a announces b, then conditional on agent 3 being type a, the outcome (a, a, b) occurs with probability 1 in stage 1 – so that a is chosen. A similar discussion applies when each player plays b with probability 1. This completes the example.

The key feature here is the *variation in the support of the distribution over player 2's types in the second stage*, depending on whether reporting is truthful or dishonest. The second stage can be reached in two ways: player 1 announces c, and the type profile reported is either $r_a = (a, a, a)$ or $r_b = (b, b, b)$. Suppose (c, r_a) occurs in the first period, so the game goes to stage 2. There are eight possible states: $\{(a, a, a), (a, a, b), \cdots, (b, b, b)\}$, according to the type of each player, (s_1, s_2, s_3). Player 2 knows the true state is in either $\mathcal{I}^a = \{(a, a, a), (a, a, b), (b, a, a), (b, a, b)\}$ or $\mathcal{I}^b = \{(a, b, a), (b, b, a), (a, b, b), (b, b, b)\}$, corresponding to whether 2 is type a or b. At information set $\mathcal{I}^a$, independent of the distribution over elements of $\mathcal{I}^a$, 2 has a dominant strategy — choose a. Similarly, at information set $\mathcal{I}^b$, 2's dominant strategy is to choose b. In the truth-telling equilibrium, (c, r_a) leads to information set $\mathcal{I}^a$ with probability 1 (where a is selected by 2), while in the deception equilibrium, (c, r_a) leads to information set $\mathcal{I}^a$ with probability μ_a and information set $\mathcal{I}^b$ with probability μ_b (where b is selected by 2). Thus, the difference between truth-telling and deception in stage 2 is that the *distribution over states changes while the action chosen at each state is the same in both cases*. At state $s = (s_1, s_2, s_3)$ the action chosen is s_2 in both truth-telling and deception, but in the truth-telling case $\text{prob}(\mathcal{I}^a \mid c, r_a) = 1$, whereas in the deception, $\text{prob}(\mathcal{I}^a \mid c, r_a) = \mu_a < 1$.

3.1.2 Example 2: Implementing a SCF through posterior induced preference variation.

In this example we illustrate how posterior distributions directly play a key role in implementing a choice rule – the change in the support of the distribution alters sequentially rational behavior, so that different behavior occurs *at the same state*, as the posterior distribution varies from truth-telling to deception.

Suppose there are three agents. Agents 1 and 2 have singleton type sets while agent three has two possible types — $S_3 = \{a, b\}$. There are four alternatives – $A = \{a, b, d, e\}$. Agents 1 and 2 have preferences that depend on the type of agent 3. The preferences of each agent i are described by a function $u_i : A \times S_3 \to \Re$. The social choice rule, $x : S_3 \to A$ is $x(a) = a$ and $x(b) = b$. Assume that for $i = 2, 3$, $u_i(a, a) = u_i(a, b) = u_i(b, a) = u_i(b, b)$. For player 2, put $u_2(d, a) > u_2(e, a)$, $u_2(e, b) > u_2(d, b)$, choosing d over e when the true state is a, and e over d when the true state is b. Letting $V_i(z) = \mu_a u_i(z, a) + \mu_b u_i(z, b)$, assume that $V_2(e) > V_2(d)$. For player 3, assume that $u_3(a, a) > u(z, c)$ for $z \in \{d, e\}$ and $c \in \{a, b\}$. Finally, for player 1, assume that $u_1(e, a) = u_1(a, a) = u_1(b, b) > u_1(e, b) > u_1(b, a) = u_1(a, b) = u_1(d, a) = u_1(d, b)$. These imply that $V_1(e) > V_1(a)$, and assume $V_1(e) > V_1(b)$. The social choice rule is implemented by the following game. In stage 1 player 3 announces a type $s_3 \in \{a, b\}$ and player 1 either challenges with an announcement $c \in \{c_a, c_b\}$ or does not, nc. If the pair chosen in stage 1 is either $\{c_a, a\}$ or $\{c_b, b\}$ the game goes to stage 2 where player 2 chooses from $\{d, e\}$. Otherwise, the game ends with the choice of 3 selected.

First, observe that truth-telling by player 3 and no challenge by 1 is an equilibrium. If player 1 challenges, c_a takes the game to stage 2 when 3 announces a and here 2 will pick d $(u_2(d, a) > u_2(e, a))$, which is worse for 1 than a $(u_1(d, a) < u_1(a, a))$. Similarly, c_b takes the game to stage 2 when 3 announces b and here 2 will pick e $(u_2(d, b) < u_2(e, b))$, which is worse for 1 than b $(u_1(e, b) < u_1(b, b))$. Next, there are three types of deception to consider: (i) both types announce a, (ii) both types announce b, and (iii) each type announces

the opposite type. We discuss these in turn. (i) If both types of 3 announce a, the challenge c_a takes the game to stage 2 where 2 picks e, since $V_2(e) > V_2(d)$. Since $V_1(e) > V_1(a)$, player 1 will select the challenge c_a. With the challenge, player 3 type b gets $u_3(e,b) < u_3(b,b)$, so player 3 type b would wish to switch to announce b. (Here, we use the fact that any sequentially consistent beliefs assign the same distribution over the types of player 3 following the challenge as is assigned by the prior distribution.) (ii) If both types of 3 announce b, the challenge c_b takes the game to stage 2 where 2 picks e, since $V_2(e) > V_2(d)$. Since $V_1(e) > V_1(b)$, player 1 will select the challenge c_b. In this case, player 3 type a gets $u_3(e,a) < u_3(b,a) = u_3(a,a)$, so player 3 type a would wish to switch to announce a. (iii) Finally, if each type of player 3 announces the opposite type, then the challenge c_a leads to choice e by player 2 when a is announced (since 3 is type b), giving player 1 utility $u_1(e,b)$. Since $u_1(e,b) > u_1(a,b)$, 1 has the incentive to challenge. In this case, player 3 type b gets $u_3(e,b)$ whereas announcing b gives player 3 type b the payoff $u_3(b,b) > u_3(e,b)$. Challenge c_b is not profitable for 1 since it produces the outcome d at state a giving utility $u_1(d,a) = u_1(b,a)$. This completes the example.

The important point in the example is the variation in behavior in period 2 of player 2, due to a change in the posterior distribution. Player 2 has just one type, *but that type plays differently depending on the posterior distribution* over three's types. To see this in terms of information sets, consider the challenge c_a with choice a by player 3 in stage 1. This takes the game to stage 2. Here, player 2 has just one information set, $\mathcal{I} = \{a,b\}$ — 2 cannot distinguish the type of player 3. If 1 challenges with c_a in the truth-telling equilibrium then if a is chosen by 3 in stage 1 the game goes to stage 2 where the posterior distribution puts probability 1 on a. Since $u_2(d,a) > u_2(e,a)$, player 2 will choose d. In the deception equilibrium where 3 chooses a independent of type, the challenge c_a takes the game to stage 2 where the posterior distribution is the same as the prior (μ_a, μ_b). Now, 2 chooses e because $V_2(e) > V_2(d)$.

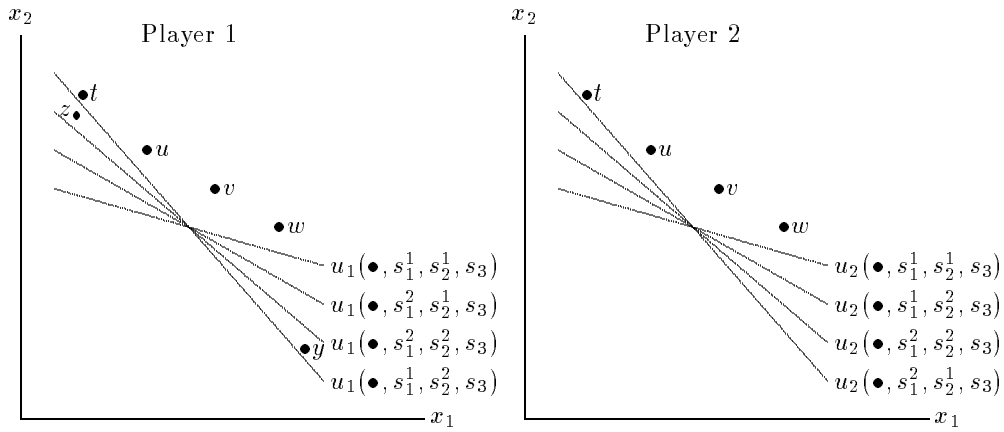
3.1.3 Example 3: Implementing a Non-Measurable SCF.

In the literature on complete information games it is known that normal form refinements such as eliminating weakly dominated strategies are more powerful than extensive form mechanisms for implementing social choice functions (Palfrey and Srivastava (1989)). Similarly, with iterative elimination of strictly dominated strategies very permissive results obtain (Abreu and Matsushima (1991)). In incomplete information games these solution criteria are also powerful. Therefore, it is somewhat surprising that in incomplete information games, there are social choice functions that cannot be implemented by either of these criteria, but can be implemented in the extensive form. This section provides an example. Rather than provide details of the extensive form, we show how a "wedge" arises in the posterior distributions (in a game where agents announce types), depending on whether strategies are truth-telling or otherwise in the first stage of a two stage game. This creates variation in the posterior distributions as type announcement strategies vary, and this is sufficient for implementation in this example. A detailed construction of an implementing game form is given later.

In this example there are three agents. Agents 1 and 2 have two types and agent three has only one type. Thus $S_1 = \{s_1^1, s_1^2\}$, $S_2 = \{s_2^1, s_2^2\}$ and $S_3 = \{s_3\}$. The joint distribution over $S_1 \times S_2$ is uniform: $\mu(s_1^i, s_2^j, s_3) = \frac{1}{4}, \quad \forall i,j \in \{1,2\}$ and $\mu(s_2^j, s_3 \mid s_1^i) = \frac{1}{2}, \quad i,j = 1,2$. The set of outcomes is a subset of $\mathcal{R}^2$, so $A \subset \mathcal{R}^2$ with representative element $a = (a_1, a_2)$. Preferences are specified as follows. For all $a \in A$,

Player 1	Player 2
$u_1(a, s_1^1, s_2^1, s_3) = 3a_1 + 9a_2$	$u_2(a, s_1^1, s_2^1, s_3) = 4a_1 + 10a_2$
$u_1(a, s_1^1, s_2^2, s_3) = 9a_1 + 3a_2$	$u_2(a, s_1^2, s_2^1, s_3) = 8a_1 + 2a_2$
$u_1(a, s_1^2, s_2^1, s_3) = 4a_1 + 8a_2$	$u_2(a, s_1^1, s_2^2, s_3) = 5a_1 + 7a_2$
$u_1(a, s_1^2, s_2^2, s_3) = 8a_1 + 4a_2$	$u_2(a, s_1^2, s_2^2, s_3) = 7a_1 + 5a_2$

And for player 3, $u_3(a, s_1^2, s_2^2, s_3) = a_1 + a_2$, $\forall s_1, s_2$.



The figure illustrates indifference curves of these preferences. Because the preferences are linear, they may be interpreted as preferences over lotteries. Thus, $V_1(a, s_1^1 \mid \mu) = 6[a_1 + a_2]$. Similarly, $V_1(a, s_1^2 \mid \mu) = 6[a_1 + a_2]$. Thus, $V_1(a, s_1^1 \mid \mu) = V_1(a, s_1^2 \mid \mu) = 6[a_1 + a_2]$, $\forall a \in A$. The function u_2 also has the property that the (interim) expected utility determined by u_2 and μ is independent of s_2: $V_2(a, s_2^1 \mid \mu) = V_1(a, s_2^2 \mid \mu) = 6[a_1 + a_2]$, $\forall a \in A$. The figure depicts indifference curves for agents 1 and 2. Thus for either player, in any game form the best response set is the same for both types, any strategy which is not weakly dominated for one type is also not weakly dominated for the other type. Because $V_1(a, s_1^1 \mid \mu) = V_1(a, s_1^2 \mid \mu), \forall a \in A$ and $V_2(a, s_2^1 \mid \mu) = V_2(a, s_2^2 \mid \mu), \forall a \in A$, the partition, $\mathcal{P}^*$, generated by iteration on equivalence classes determined by these functions is the coarsest partition: $\mathcal{P}^* = \{S_1\} \times \{S_2\} \times \{S_3\}$. The only functions measurable relative to this information are constant functions.

Let $\{t, u, v, w\}$ be four points in $A \subset \mathcal{R}^2$. (There is no difficulty concerning the existence of incentive compatible allocations. For example, $z_1 + z_2 = k, \forall z \in \{t, u, v, w\}$, then all allocation values lie on an indifference curve of the (type independent) interim expected payoff and hence satisfy incentive compatibility).

Now, assume that $\exists \bar{a} \in A$, $\bar{a}_i > 0, i = 1, 2$, such that $a \in A$, if $0 \leq a_i \leq \bar{a}_i, i = 1, 2$. Consider the social choice function x defined: $x(s_1^1, s_2^1, s_3) = (t, t, \bar{a} - 2t)$, $x(s_1^2, s_2^1, s_3) = (u, u, \bar{a} - 2u)$, $x(s_1^1, s_2^2, s_3) = (v, v, \bar{a} - 2v)$, $x(s_1^2, s_2^2, s_3) = (w, w, \bar{a} - 2w)$. This is interpreted as follows: at state (s_1^1, s_2^1, s_3), the allocation to agents is $(t, t, \bar{a} - 2t)$, with a similar interpretation applying to the other states. Finally, assume that $\forall z \in \{t, u, v, w\}$, $\bar{a}_i - 2z_i > 0, i = 1, 2$.

Since this function varies over S it is not measurable with respect to $\mathcal{P}^*$. Now, consider a two stage game where in the first period agents announce types. Whether or not the game goes to stage two depends on some action of the third player. Because the third player has just one type, all consistent distributions over the types of players 1 and 2 are determined by Bayes' rule. Consider a deception $\alpha = (\alpha_1, \alpha_2)$ where $\alpha_1 = \hat{\alpha}_1$ (truth-telling), while $\alpha_2(s_2^1) = s_2^2$, $\alpha_2(s_2^2) = s_2^1$. Let $s^* = (s_1^1, s_2^2)$. Now consider two possibilities. In the first, both agents play the truth-telling strategy $(\hat{\alpha}) = (\hat{\alpha}_1, \hat{\alpha}_2)$. In this case, if s^* is announced (and this event has positive probability), then if the game goes to stage two, the posterior over types puts probability 1 on s^* (since agents are using the truth-telling strategy $\hat{\alpha}$. Then, if agent 1 is faced with a choice between y and z, with probability 1 agent 1 will choose y.

Now, suppose that the deception α is played and s^* is announced. If the game goes to stage two, then the posterior distribution puts probability 1 on $\tilde{s} = (s_1^1, s_2^1)$. If agent 1 is faced with the choice between y and z, with probability 1 agent 1 will choose z. This variation in behavior is all that is required at stage two for extensive form implementation. Other possible deceptions are treated similarly. For any α an appropriate (y, z) pair exist which yields a preference switch. A general mechanism that implements this social choice rule is given later.

Remark 4 The mechanism we give implements the social choice function in pure strategies. If mixed strategies are allowed, and we permit randomization in deceptions, the social choice rule in this example is still implementable. The reason is simple. Even when randomization is allowed, under any deception the posterior distribution at some type announcement will necessarily be different from the posterior distribution under truth-telling at that announcement. For example, if agent 2, type s_2^1 announces s_2^2 with positive probability, then the posterior distribution on two's types following signal s_2^2 puts positive probability on type s_2^1, whereas in truth-telling, the signal s_2^2 leads to a posterior which puts probability 1 on s_2^2. Detailed computations are given in Bergin and Sen (1993).

4 Sufficient Conditions for Implementation of a SCF.

Below, we introduce *posterior reversal*, a condition that relates to the reversal of ranking of outcomes by some agent at different posterior distributions. Before giving the condition it is necessary to introduce some terminology relating to belief systems in extensive form games.

4.1 Belief Systems.

Let $\alpha_i : S_i \to \Delta(S_i)$, $i = 1, \ldots, n$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\alpha : S \to \times_i \Delta(S_i)$. Thus, α^i associates a probability distribution on S_i to each $s_i \in S_i$. Write $\alpha_i^{s_i}$ to denote the distribution on S_i at s_i, so $\alpha_i^{s_i}(\tilde{s}_i)$ denotes the probability of $\tilde{s}_i$ under the distribution $\alpha_i^{s_i}$. Given the prior distribution μ and a deception α, the probability of $\tilde{s}$ is $\mu^\alpha(\tilde{s}) = \sum_{s \in S} [\times_{i=1}^n \alpha_i^{s_i}(\tilde{s}_i)]\mu(s)$.

Provided $\mu^\alpha(\tilde{s}) > 0$, define the (posterior) distribution:

$$\mu^\alpha(s \mid \tilde{s}) = \frac{[\times_{i=1}^n \alpha_i^{s_i}(\tilde{s}_i)]\mu(s)}{\sum_{\hat{s} \in S} [\times_{i=1}^n \alpha_i^{\hat{s}_i}(\tilde{s}_i)]\mu(\hat{s})}, \ \forall s \in S.$$

If $\mu^\alpha(\tilde{s}) = 0$, the definition is arbitrary.

Different equilibrium refinements use different means of restricting beliefs off the equilibrium path, but all beliefs refinements satisfy Bayes' rule whenever it is applicable. The case of primary interest is where an event has 0 probability because of a deviation by just one agent. In the present context, this case occurs when $\mu(s) > 0$, $[\times_{j \neq i} \alpha_j^{s_j}(\tilde{s}_j)] > 0$ and $\alpha_i^{s_i}(\tilde{s}_i) = 0, \forall s_i \in S_i$. Thus $\tilde{s}$ has 0 probability because of i ($\text{Prob}(\tilde{s}_{-i}) > 0$ and $\text{Prob}(\tilde{s}_{-i}, \tilde{s}_i) = 0$). In this case, given a theory for defining beliefs along paths that have zero probability, denote the set of posterior distributions consistent with the theory by $\tilde{s}$ by $C(\tilde{s}, i, \alpha)$. (For example, sequentially consistent beliefs of sequential equilibrium are obtained by taking any sequence $\{\alpha^n\}$ that is fully mixed, so $\alpha_i^n : S_i \to \text{interior}(\Delta(S_i))$), with $\{\alpha^n\} \to \alpha$. Taking any limit of the associated belief sequence gives a consistent beliefs system.)

If all agents are playing truth-telling strategies, so that for each j the reporting strategy is $\hat{\alpha}^j$, $\hat{\alpha}_j^{s_j}(\tilde{s}_j) = 1$, if $s_j = \tilde{s}_j$ and 0 otherwise. Or, letting $\delta_j^{s_j}$ denote the distribution with unit mass at s_j, $\hat{\alpha}_j^{s_j} = \delta_j^{s_j}$. Then $\text{Prob}(\tilde{s}_{-i} \mid \tilde{s}_{-i}) = 1$. In this case, the set of consistent distributions has the form:

$$C(\tilde{s}, i, \alpha) = [\times_{j \neq i} \delta_j^{\tilde{s}_j}] \times \Delta(S_i).$$

Next, suppose that for some $j \neq i$, there is a pair $s_j, \tilde{s}_j$ with $s_j \neq \tilde{s}_j$ and $\alpha_j^{s_j}(\tilde{s}_j) > 0$. Pick s_k and $\tilde{s}_k$, $k \neq i, j$ such that $\alpha_k^{s_k}(\tilde{s}_k) > 0, \forall k \neq i, j$. Thus, $s_{-i} \neq \tilde{s}_{-i}$ and $[\times_{i \neq j} \alpha_i^{s_i}(\tilde{s}_i)] > 0$, so that $\text{Prob}(s_{-i} \mid \tilde{s}_{-i}) > 0$. So, if $\hat{\mu}$ is in $C(\tilde{s}, i, \alpha)$, then $\sum_{s_i} \hat{\mu}(s_{-i}, s_i) = \hat{\mu}(s_{-i}) \geq \gamma > 0$, where the lower bound γ depends only on $[\times_{j \neq i} \alpha_j^{s_j}(\tilde{s}_j)]$ and the prior distribution μ. Thus, there is some s_i such that $\hat{\mu}(s_{-i}, s_i) \geq [1/\#S_i]\gamma = \bar{\gamma} > 0$. Recall that the set of consistent distributions under truth-telling is $C(\tilde{s}, i, \alpha) = [\times_{j \neq i} \delta_j^{\tilde{s}_j}] \times \Delta(S_i)$. Thus, if $\hat{\mu}$ is a posterior distribution on S in the consistent set under truth-telling, then given the observation of $\tilde{s}$, with $s_{-i} \neq \tilde{s}_{-i}$, $\hat{\mu}(s_{-i}, s_i) = 0, \forall s_i \in S_i$, while under the deception α, $\hat{\mu}(s_{-i}, s_i) > 0$, for some s_i. Summarizing, (s_{-i}, s_i) is a point in type space which in any consistent distribution under the deception has probability greater than or equal to $\bar{\gamma}$ (> 0) and in any consistent distribution under truth-telling has probability 0. (In sequential equilibrium out of equilibrium beliefs are generated as the limits of posterior distributions determined by fully mixed strategies. A detailed discussion of belief systems in the present context is given in Bergin and Sen (1993).)

4.2 Posterior Reversal.

The next condition is central to our sufficiency result. The notation $supp_j \mu$ denotes the support of the distribution in S_j (the support of the marginal of μ on S_j).

Definition 9 *The social choice function x satisfies* **posterior reversal** *if for each $\alpha \in D \setminus \{\hat{\alpha}\}$, $x \neq x_\alpha$, $\exists\, i, j \in I$, $\bar{s} \in \alpha(S)$ and constant allocations $\bar{y}, \bar{z} \in X$ such that*

1. $\forall\, \alpha_i \in D_i,\ s_i \in S_i$

$$x R^i(s_i, \mu) y_{(\alpha_i, \hat{\alpha}_{-i})} \quad \text{where } y(\bar{s}) = \bar{y},\ y(s') = x(s'), s' \neq \bar{s}$$

2. $\exists \mu' \in C(\bar{s}, i, \hat{\alpha})$ *such that* $\forall\, s_j \in supp_j \mu'$, $j \neq i$, *and* $\forall s_j \in S_j$ *if* $j = i$

$$\bar{y} R^j(s_j, \mu') \bar{z}.$$

3. $\forall\, \mu' \in C(\bar{s}, i, \alpha)$, $\exists\, t_j \in supp_j \mu'$, *if* $j \neq i$, *and* $\exists t_j \in S_j$ *if* $j = i$, *such that*

$$\bar{z} P^j(t_j, \mu') \bar{y}.$$

4. $\exists a^* \in A$, $s_i \in S_i$,

$$\sum_{\{s_{-i} \mid (s_{-i}, s_i) \in \alpha^{-1}(\bar{s})\}} u_i(a^*, s) \mu(s_{-i} \mid s_i) > \sum_{\{s_{-i} \mid (s_{-i}, s_i) \in \alpha^{-1}(\bar{s})\}} u_i(x(\bar{s}), s) \mu(s_{-i} \mid s_i),$$

As with Bayesian monotonicity we say that the social choice rule satisfies posterior reversal at α when these conditions are satisfied at that α. Because posterior reversal imposes conditions on preferences at posterior distributions, Bayesian monotonicity is not a special case of posterior reversal. (We consider how posterior reversal relates to other concepts in section 4.4.) Recall that the consistent belief sets $\mathcal{C}(s, i, \alpha)$ can be calculated solely with knowledge of the sets of player types and the prior distribution. Thus, posterior reversal can be checked without reference to a specific game form. Intuitively, in truth-telling, if the game goes to stage 2, then $\bar{y}$ is supported as an equilibrium (condition 2). Therefore the challenge (by i) in stage 1 is not profitable (condition 1). In the deception, $\bar{y}$ is not supported in stage 2 and the switch (by j) to

alternative $\bar{z}$ forces an equilibrium switch to outcome a^*, desired by the challenger (condition 4) (in the game we construct, the switch from $\bar{y}$ to $\bar{z}$ is used to make player i a dictator in stage 2). Posterior reversal is most easily explained in terms of the examples given earlier. However, before discussing them, some general remarks are appropriate.

Remark 5 Observe that conditions 2 and 3 of posterior reversal imply that $\exists \hat{\mu} \in C(\bar{s}, i, \hat{\alpha})$, $\{\hat{\mu}\} \notin C(\bar{s}, i, \alpha)$. In the context of an extensive form game, the interpretation of this requirement is straightforward. In truth-telling we wish to support some equilibrium outcome with equilibrium behavior on the subform reached by signal, $\bar{s}$. If the truth-telling posterior were also in the set of consistent beliefs under the deception, then the equilibrium behavior at that subform in truth-telling is also equilibrium behavior under the deception at that signal. In this case, the signal triggers the same outcome under truth-telling as deception and the extensive form adds nothing that is not achievable in a normal form game.

Remark 6 It is worth noting that for any deception $\alpha \neq \hat{\alpha}$ the sequential consistency condition of sequential equilibrium implies that there is some $\bar{s}$ such that $C(\bar{s}, i, \hat{\alpha}) \cap C(\bar{s}, i, \alpha) = \emptyset$. Furthermore, this observation is valid whether or not randomization is allowed. The reason is simple: in a deception, some type of some player is announcing another type with positive probability. For example, suppose that s_i announces $\tilde{s}_i$ with positive probability. Then, the conditional probability on s_i given $\tilde{s}_i$ is positive, whereas in truth-telling the signal $\tilde{s}_i$ implies the true type is $\tilde{s}_i$ with probability 1. This observation also applies to most forms of perfect Bayesian equilibrium.

Remark 7 In a general extensive form game agents' strategy choices at any point may be much larger than their type space. Thus, the uncertainty facing a player at any point in the game may be not just the types of other players, but also (for example) previous choices made. Therefore, in general, the consistent distributions may be defined on a larger space than the set of types. Nevertheless, given any social choice rule that fails Bayesian monotonicity at some deception α, if the rule is implemented in an extensive form game with one round of signaling, then it is necessarily the case that at some subform of the game the beliefs in the "truth-telling" equilibrium are disjoint from the beliefs implied by the candidate deception.

Example 1 (Voting). Consider the deception $\alpha(s) = (a, a, a)$, $\forall s \in S$ in the majority voting model. Observe that in truth-telling, the posterior distribution puts probability 1 on player 2 being type a, so type a is the only player in the support of the truth-telling distribution (in truth-telling $\mu' \in C(\bar{s}, i, \hat{\alpha})$ implies that $supp_2 \mu' = \{a\}$). In the posterior reversal condition, take $i = 1$, $j = 2$, let $\bar{s} = (a, a, a)$, $\bar{y} = a$ and $\bar{z} = b$. Let s_j in condition 2 be a and let t_j in condition 3 be b. Since type a prefers a over b, condition 2 is satisfied — every type of player 2 in the support of the truth-telling posterior ranks a at least as good as b. Condition 1 is satisfied because $y(a, a, a) = a = x(a, a, a)$, so that x and y are identical. For condition 3, in the deception the posterior distribution coincides with the prior, so that b is in the support of the distribution over two's types. Since type b of player 2 strictly prefers b to a ($\bar{z}$ to $\bar{y}$), condition 3 is satisfied. Finally, consider condition 4. Pick $a^* = b \in A$ and put $s_1 = b$. Since $\alpha^{-1}(\bar{s}) = S$, the left side of inequality 4 is (recall preferences in the example satisfy private values): $\sum_{s_{-1} \in S_{-1}} u_i(b, b)\mu(s_{-1} \mid b) = u_i(b, b)$ and the right-hand side reduces to $u_i(a, b)$. Since $u_i(b, b) > u_i(a, b)$, condition 4 is satisfied.

Example 2 (Type preference variation). Consider the deception where both types of player 3 announce a. Put $\bar{s} = a$ and set $\bar{y} = d$, $\bar{z} = e$. Thus, $y(\bar{s}) = d$ and $y(b) = x(b) = b$. In the posterior reversal condition put $i = 1$ and $j = 2$. Condition 1 is satisfied since $u_1(a, a) > u_1(d, a)$. To see that condition 2 is satisfied note that the truth-telling posterior distribution puts probability 1 on 3 being type 1. Player 2 has just one type

so the support condition just requires that the optimal choice for player two given this posterior distribution is to choose d. Since $u_2(d,a) > u_2(e,a)$, this is satisfied. Under the deception the posterior coincides with the prior and then the optimal choice for 2 is e since $V_2(e) > V_2(d)$, so that condition 3 is satisfied. Finally, condition 4 is satisfied since for player 1, $V_1(e) = u_1(e,a)\mu_a + u_1(e,b)\mu_b > u_1(a,a)\mu_a + u_1(a,b)\mu_b = V_1(b)$.

The third example is similar and will not be discussed. The main result of this section is:

Theorem 1 *Let $n \geq 3$. Suppose that the SCF, $F = \{x\}$ satisfies self-selection, posterior reversal at any α which fails Bayesian monotonicity, and economic environments. Then F is implementable in an extensive form game in sequential equilibrium.*

Proof: The implementing game form is given in the appendix. ■

Remark 8 Any equilibrium concept that satisfies sequential rationality and which in a direct reporting mechanism generates beliefs (where beliefs are computed conditional on the report) at every information set such that (PR) holds, is adequate. We use sequential equilibrium.

Posterior reversal may be relaxed in two directions: (a) in challenging a given α, many $\bar{s}$'s may take the game to stage 2 (so that any report in a subset, $S_\alpha \subset \alpha(S)$ may take the game to the second stage), and (b) instead of constant allocations in stage 2, type dependent allocations may be allowed.

4.3 Distribution-Free Cases of Posterior Reversal.

It may sometimes be useful to have conditions which imply posterior reversal but are not phrased in terms of posterior distributions. In what follows we give two sets of conditions which imply posterior reversal in different circumstances. In the case of private values the conditions are relatively simple.

Theorem 2 *Posterior reversal holds if the following conditions are satisfied. Taking each α and associated triple $i, j, \bar{s}$ and constant outcomes $\bar{y}, \bar{z}$.*

1. $[u_i(x(\bar{s}), \bar{s}_{-i}, s_i) - u_i(\bar{y}, \bar{s}_{-i}, s_i)] \geq 0, \forall s_i \in S_i$
2. *(2.1) if $i \neq j$,* $u_j(\bar{y}, \bar{s}_{-i}, s_i^*) \geq u_j(\bar{z}, \bar{s}_{-i}, s_i^*)$, *some* $s_i^* \in S_i$
   (2.2) if $i = j$, $u_i(\bar{y}, \bar{s}_{-i}, s_i) \geq u_i(\bar{z}, \bar{s}_{-i}, s_i)$, $\forall s_i \in S_i$
3. *(3.1) if $i \neq j, \exists t_j \in \alpha_j^{-1}(\bar{s}_j)$,*

$$u_j(\bar{z}, s_j, t_j) > u_j(\bar{y}, s_{-j}, t_j), \forall s_{-j} \in \varphi(\alpha, \bar{s}, \{i,j\}) \times S_i.$$

   (3.2) if $i = j, \exists t_i \in S_i$, $u_i(\bar{z}, s) > u_i(\bar{y}, s), \forall s \in \varphi(\alpha, \bar{s}, i) \times \{t_i\}$.
4. $\exists a^* \in A$, $s_i \in S_i$,

$$\sum_{\{s_{-i} | (s_{-i}, s_i) \in \alpha^{-1}(\bar{s})\}} u_i(a^*, s)\mu(s_{-i} \mid s_i) > \sum_{\{s_{-i} | (s_{-i}, s_i) \in \alpha^{-1}(\bar{s})\}} u_i(x(\bar{s}), s)\mu(s_{-i} \mid s_i),$$

Proof: The proof is given in the appendix. ■

In point 3 of the theorem, φ is defined as follows. Given $I' \subset I$, let $\varphi(\alpha, \bar{s}, I') = \{s_k \in S_k \mid k \notin I', \alpha_k(s_k) = \bar{s}_k\}$. This identifies the set of agents *not* in I' whose types could be mapped under α to $\bar{s}$. In particular, if $\mu' \in \mathcal{C}(\bar{s}, i, \alpha)$, then μ' has support in $\varphi(\alpha, \bar{s}, i) \times S_i$.

A special case is private values: for each i, $u_i(a, s_{-i}, s_i) = u_i(a, s_i)$, $\forall a \in A, s_{-i} \in S_{-i}, s_i \in S_i$. In this case, theorem 2 specializes further.

Theorem 3 *Suppose that preferences satisfy private values. Then with $\alpha, i, j, \bar{s}, \bar{y}$ and $\bar{z}$ as in theorem 2, posterior reversal condition is satisfied if and only if $i \neq j$ and:*

1. $u_i(x(\bar{s}), s_i) \geq u_i(y, s_i), \forall s_i \in S_i$
2. $u_j(y, \bar{s}_j) \geq u_j(z, \bar{s}_j)$
3. $\exists t_j \in \alpha_j^{-1}(\bar{s}_j),\ u_j(z, t_j) > u_j(y, t_j)$
4. $\exists a^* \in A,\ s_i \in S_i,\ u_i(a^*, s_i) > u_i(x(\bar{s}), s_i)$

Proof: The proof follows directly from theorem 2 with the modified utility function. ■

Note that in this case i cannot equal j. (In condition 2, the specialization would give (for $i = j$), $u_i(y, s_i) \geq u_i(z, s_i), \forall s_i$, while condition 3 would require that $\exists t_i \in S_i$ such that $u_i(z, s_i) > u_i(y, s_i)$.) Condition 1 requires that given a deception α, there is some $\bar{s} \in \alpha(S)$, such that $x(\bar{s})$ is (weakly) preferred by all types of player i (condition 1) to y. Condition 2 requires that player j, type $\bar{s}_j$ (weakly) prefer y to z, while condition 3 requires that some type of player j reporting $\bar{s}_j$ in the deception (some t_j such that $\alpha(t_j) = \bar{s}_j$), strictly prefer z to y. Finally, condition 4 requires that there is some type of player i for which $x(\bar{s})$ is not top ranked.

4.4 Posterior Reversal, Bayesian monotonicity and related conditions.

Posterior reversal emphasizes the role of belief variation between truth telling and deception to eliminate candidate deception equilibria. This contrasts with Bayesian monotonicity which emphasizes the comparison between alternative outcomes under truth telling and deception. In this section we develop a connection between Bayesian monotonicity and posterior reversal.

When multistage games are considered, players reports in earlier stages may influence subsequent behavior. Here, we focus on the case where there is one round of reports. Given a deception α and a function $f : S \times S$ to A, define $f_\alpha(s', s) \equiv f(\alpha(s'), s)$. Interpret $f(s', s)$ as the outcome when there is an initial report s' and the true type profile is s. Thus $f(s, s)$ is the value of f on the diagonal of $S \times S$ at (s, s) and $f_\alpha(s, s)$ is the value of f at $(\alpha(s), s)$, and gives the outcome at type profile s when the deception α is played. Also, $f_{\alpha_i}(s, s) = f((\alpha_i(s_i), s_{-i}), s)$, so that the notation parallels standard usage. If α is not the identity function, then for some s', $\alpha(s') \neq s'$. Define $V_i(f, s_i \mid \mu) = \sum_{s_{-i} \in S_{-i}} u_i(f(s, s), s)\mu(s_{-i} \mid s_i)$. Write $fR(s_i, \mu)g$ if $V_i(f, s_i \mid \mu) \geq V_i(g, s_i \mid \mu)$ and $fP(s_i, \mu)g$ if the inequality is strict. If the function $g : S \to A$, then $V_i(g, s_i \mid \mu)$ is defined as earlier (following definition 1).

Posterior reversal utilizes variation of beliefs on a subform (comparing truth-telling and deception) to generate preference reversals. As mentioned earlier, there are two distinct ways in which this can occur. The first occurs when stage two behavior doesn't change with beliefs, but the change in beliefs alters the distribution over outcomes and generates the preference reversal. We shall call this *generalized Bayesian monotonicity*. The second case arises when variation in beliefs at stage two leads to a change in behavior by specific types of some players. We shall call this *chain reversal*. Next, we give formal definitions and relate them to the earlier examples.

Definition 10 *A social choice rule, x, satisfies* **generalized Bayesian monotonicity** *(GBM) if given $\alpha \in D$, $x_\alpha \neq x$, $\exists i \in I, s_i \in S_i$ and $y : S \times S \to A$ such that:*

a. $xR^i(t_i, \mu)y_{\alpha_i(s_i)},\ \forall t_i \in S_i$

b. $y_\alpha P^i(s_i, \mu)x_\alpha$

Note that when y in the GBM condition is independent of the second argument, GBM reduces to Bayesian monotonicity. Intuitively, the first argument of $y(s', s)$, s', corresponds to the reported type and the second

argument corresponds to the "true" type profile. GBM arises when at the relevant subform, each player uses the same strategy (on that subform) in both the truth-telling equilibrium and the candidate deception equilibrium. However, with "truth telling", the report $\hat{s}$ means that subsequent stage choices are made by the types in $\hat{s}$; whereas in a deception with $s' \neq \hat{s}$ reporting $\hat{s}$, a report of $\hat{s}$ in stage 1 means that the choices made in subsequent stages are with positive probability made by s'. The voting example given earlier illustrates GBM.

Example 1. Consider the deception $\alpha(s) = (a, a, a), \forall s \in S$. Define $y(a, a, a, s_1, s_2, s_3) = y(b, a, a, s_1, s_2, s_3) = s_2$ and $y(\hat{s_1}, \hat{s}_2, \hat{s}_3, s_1, s_2, s_3) = x(a, \hat{s}_2, \hat{s}_3)$, if either $s_2 \neq a$ or $s_3 \neq a$. Let $i = 1$ and $s_i = b$ in the definition of GBM.

In general, changes in beliefs will alter the behavior of given types; the same type may act differently as beliefs change. This is captured by chain reversal.

Definition 11 *A social choice rule, x, satisfies* **chain reversal** *(CR) if given $\alpha \in D$, $x_\alpha \neq x$, $\exists i, j \in I$, $s_i \in S_i$, $s_j \in S_j$ and functions y and z, $y : S \times S \to A$, $z : S \times S \to A$,*

a. $xR^i(t_i, \mu)y_{\alpha_i(s_i)}$, $\forall t_i \in S_i$,

b. $y_{\alpha_i(s_i)}R^j(t_j, \mu)z_{\alpha_i(s_i)}$, $\forall t_j \in S_j$

c. $z_\alpha P^j(s_j, \mu)y_\alpha$

Here, in truth telling, x is supported by y — a challenge by i leads to y which is no better than x for any type of i; and y is preferable to some z by all types of player j so that y is "supported by y thus deterring a challenge by i. In deception, some type of j has a preference flip, choosing z over y under the deception (z_α preferred to y_α). The preference reversal of j leads to different choices by j and may be used to create differential incentives for i, comparing truth-telling and deception — hence the term "chain reversal". Chain reversal is illustrated by example 2.

Example 2. Consider the deception $\alpha(s_3) = a, \forall s_3 \in S_3$. Let $y(a, a) = y(a, b) = d$ and $y(b, a) = y(b, b) = b$. Also, let $e = z(a, a) = z_\alpha(a, a) = z(a, b) = z_\alpha(b, b)$ and put $z(b, a) = z(b, b) = b$.

We conclude this section by relating these concepts to games with one round of signaling. Call an incomplete information extensive form game a *game with one round of signaling*, if all players move simultaneously in a first stage, and there are no equilibria where later stages are reached with positive probability.[3] Thus, player i has a stage 1 message space C_i, and the set of possible stage 1 messages, $C = \times_{i=1}^n C_i$, is partitioned into two sets C^1 and C^2. A message $c \in C^1$ terminates the game with some outcome $g(c) \in A$, while a message $c \in C^2$ leads to a subsequent stage. A Strategy for i, (σ_i, τ_i) associates a choice at each information set: the first stage component of i's strategy has the form $\sigma_i : S_i \to C_i$, and the second stage component $\tau_i : S_i \times C$, (where τ_i is constant on all c's in any given information set of i.) By assumption, if (σ, τ) is an equilibrium, then for all $s \in S$ $\sigma(s) \in C^1$. (And τ is assumed to satisfy sequential rationality at every information set.)

[3] Recently, Brusco (1997) has provided an example of social choice rule that is not implementable in a game with one round of signaling. In the example, what a player knows about other players varies with that player's type; and the solution concept imposes the "non-expanding support" belief restriction. In this paper, we assume that each player only knows their own type: player i type s_i knows only that the true state is $\{s_i\} \times S_{-i}$, a situation that we may define as "private information". The issue of how the support of the distribution varies along subforms doesn't arise here because we start with full support and with one round of signaling the way in which beliefs are determined beyond the second stage is not relevant.

Proposition 1 *Let Γ be a game with one round of signaling which implements x, with implementing strategy $(\bar{\sigma}, \bar{\tau})$. Suppose that for some system of (stage two) beliefs consistent with $\bar{\sigma} \circ \alpha$ there is some $\tilde{\tau}$ yielding an equilibrium on every subform of the game (in stage 2).*[4] *Then:*

a. If, under a belief system determined by $\bar{\sigma} \circ \alpha$, $\bar{\tau}$ remains an equilibrium across subforms (reachable by a deviation of just one player from $\bar{\sigma} \circ \alpha$), then x satisfies GBM at α.

b. If, under any belief system determined by $\bar{\sigma} \circ \alpha$, $\bar{\tau}$ does not define an equilibrium on each subform (reachable by a deviation of just one player from $\bar{\sigma} \circ \alpha$), then x satisfies CR.

(See Bergin and Sen (1997) for a proof.)

Remark 9 Suppose that the social choice rule fails Bayesian monotonicity at some α but is implemented by a game with one round of signaling. Suppose that in the implementing game, b. of proposition 1 holds, so that chain reversal applies. Then, with player i, type s_i as in the definition of chain reversal, $\exists w : S \times S \to A$ such that $w_\alpha P^i(s_i, \mu) x_\alpha$. Thus, for player i one obtains $x R^i(t_i, \mu) y_{\alpha(s_i)}$, $\forall t_i \in S_i$ and $w_\alpha P^i(s_i, \mu) x_\alpha$. The interpretation is that in the deception i type s_i gets w_α with the elimination of y_α by j. In example 2 above, define $w = z$.)

5 Related Literature.

Baliga (1993) and Brusco (1995) develop general necessary and sufficient conditions for extensive form implementation in incomplete information environments. Both consider extensive form mechanisms with an arbitrary but finite number of stages. Baliga considers the case of private values and independent types, using perfect sequential equilibrium as the solution concept. This refines sequential equilibrium with forward induction requirements. Brusco allows correlated types and individuals preferences may depend on the types of others; perfect Bayesian equilibrium is used as the solution concept. In the construction of our implementing game form, we use sequential equilibrium. Thus, in terms of solution concepts, the one used by Baliga is a refinement of the concept we use, and this in turn refines the solution concept use by Brusco. The use of different solution concepts is not unimportant (especially if trying to narrow the game between necessary and sufficient conditions for extensive form implementation.) In all three papers, candidate "deception" equilibria are "knocked out" by play off the equilibrium path. Thus, for example, for a given game form, using perfect Bayesian equilibrium (instead of sequential equilibrium) makes it is easier to support "truth-telling" as an equilibrium since there is a larger set of beliefs to support continuation payoffs along a path reached by deviation; on the other hand it is more difficult to eliminate deception equilibria — for the same reason — because there is a larger set of equilibrium continuation payoffs available to support the deception as an equilibrium. The converse applies when perfect sequential equilibrium is used.

In the works of Brusco and Baliga, the necessary conditions parallel the complete information conditions given in Abreu and Sen (1990) and Moore and Repullo (1988). In both, the necessary conditions require a string of social choice rules and associated beliefs, such that at the end of the string, a preference reversal occurs (when expected utility is computed with the posterior distribution defined under truth telling vis-a-vis the expected utility computed with the posterior distribution defined under deception.) Their sufficiency

[4] One other possibility, raised by a referee, is that the game might have no equilibrium on some the subform (with beliefs determined relative to $\bar{\sigma} \circ \alpha$.) Then there is a "preference-flip" — such as player j in the chain reversal condition — but no direct connection back to the first stage. This is particularly clear in the case where the subform is reached by a correlated deviation in stage 1 (relative to $\bar{\sigma}$ the first component of the implementing equilibrium strategy). Here, the precise way in which beliefs are constructed is critical to establishing a connection back to stage 1.

conditions build on the necessary conditions, and are somewhat complex. Since the sets of consistent or admissible beliefs vary with the solution concept, the circumstances under which a preference reversal occurs will also.[5]

Our sufficiency condition starts from the same fundamental requirement — that somewhere a preference reversal occurs in the extensive form, comparing truth-telling to deception. As mentioned above, in both Baliga's and Brusco's work, this observation forms the basis for the necessary conditions. On the other hand, we take a different direction and pursue further the way in which a reversal occurs; identifying "splitting" properties of posterior distributions comparing truth-telling to deception, and identifying the exact ways in which variations in beliefs translate into variations in the distributions over outcomes. A key central observation in our work is that somewhere in the extensive form, at some subform there must be some belief determined in truth-telling which is disjoint from the set of possible beliefs under deception. In our framework this then leads to the identification of conditions under which this belief separation can be translated into a preference reversal. (Regardless of the solution concept use, this belief separation must occur somewhere, so the basic idea is applicable, whatever solution one has in mind.) This viewpoint lends itself naturally to a signaling interpretation: a given signal generates different beliefs under truth-telling than under deception. For many problems of mechanism design in incomplete information environments, this signaling viewpoint provides a useful approach.

Finally, some comments on the use of different solution concepts and the "gap" between necessary conditions and sufficiency conditions are appropriate. Consider a multistage game. The possibility exists that a candidate deception equilibrium is eliminated by the non-existence of equilibrium at a subform which is unreachable by a unilateral deviation of any player. At first sight, this would imply that it may not be necessary to have a connection back to the first stage (in terms of preference reversals) to eliminate a candidate deception equilibrium. When perfect Bayesian equilibrium is the solution concept, beliefs on a subform reached by a correlated deviation are a superset of the beliefs determined by a unilateral deviation — comparing subforms at the same stage that are identical, apart from assigned beliefs. (Furthermore, beliefs are determined by local considerations at each of the subforms: the assignment of beliefs at one subform does not restrict the assignment of beliefs at the other.) As a consequence, in perfect Bayesian equilibrium, nonexistence of equilibrium for any beliefs at the subform reached with a correlated deviation implies nonexistence under all beliefs at the subform reached by the unilateral deviation (when identical subforms, apart from beliefs, are considered). This in turn is used to provide the chain leading to a preference reversal in the necessary conditions of Brusco. (With independent types (as in the work of Baliga), the issue may not arise.) In contrast, with sequential equilibrium, beliefs are assigned simultaneously with one test sequence. The issue of whether or not one can identify beliefs at certain subforms as subsets of beliefs at another is then central in determining whether or not the preference reversal in the extensive form must be connected back to the first stage through a chain of preferences.

[5] With many stages, the issues become more complex. For example, if the sequential consistency condition of sequential equilibrium is used to determine distributions on subforms, then posteriors under multiple deviations quickly become non-informative (recall that with just one deviation in truth telling, the "off the equilibrium path" posterior had the form $\mathcal{C}(\tilde{s}, i, \alpha) = [\times_{j \neq i} \delta_j^{\tilde{s}_j}] \times \Delta(S_i)$.) Use of the non-expanding support condition of perfect sequential equilibrium can alleviate this problem; but at least with sequential equilibrium the benefits from having additional stages is mixed.

6 Appendix.

6.1 Proof of Theorem 1.

We describe a game form and then confirm that it implements the social choice function. We assume that at each α, posterior reversal is satisfied. When Bayesian monotonicity is satisfied at α, we can build this into the first stage of the game directly, and there is no need to exploit posterior reversal at all.

A. The Game Form

If posterior reversality is satisfied, then a unique $(i, j, \bar{s}, y, z)$ may be associated with each $\alpha \in D \setminus \{\hat{\alpha}\}$. Identify these respectively by $i(\alpha), j(\alpha), s(\alpha), y^{\alpha}$ and z^{α}. Set $D(i) = \{\alpha \in D \setminus \{\hat{\alpha}\} \mid i(\alpha) = i\}$.

The game has two stages: 1 and 2. In stage 1, agent i selects (conditional on type) an element of the set $M_i = S_i \times D \times \mathcal{N}$ where $\mathcal{N}$ is the set of non-negative integers. Thus, agent i "announces" a type, $s_i \in S_i$, a deception $\alpha(i) \in D$, an integer $n_i \in \mathcal{N}$.

A detailed description of the sequencing of events is as follows.

Stage 1

(1) If either

(a) $\#\{k \in I | n_k = 0\} \leq n - 2$ or

(b) $\exists\ i, j \in I$ such that $\alpha(i) \neq \hat{\alpha} \neq \alpha(j)$, then a unique player, i^* is identified: $i^* = \min \arg\max_i\ n_i\}$. This player (a dictator) selects an outcome in A.

(2) If $\exists i \in I$ such that

1. $\forall\ j \neq i, \alpha(j) = \hat{\alpha}, \alpha(i) = \alpha \neq \hat{\alpha}$ and $\alpha \in D(i)$
2. $\#\{k \in I | n_k = 0\} = n$.
3. $s_k = \bar{s}_k\ \forall\ k \in I$ where $\bar{s} = s(\alpha)$.

then the game proceeds to Stage 2.

(3) In all other cases, the outcome is $x(s_1, s_2, ..., s_n)$.

Thus, an "integer game" is played if more than one agent announces a non-zero integer or if there are at least two agents who announce non-identity deceptions. The game moves to Stage 2 if and only if (a) all agents pick zero (b) all but one agent announces $\hat{\alpha}$ with the remaining agent, say i announcing some $\alpha \in D(i), \alpha \neq \hat{\alpha}$ and (c) $\bar{s} = s(\alpha)$ is the vector of types announced by agents.

In the game, the message space in stage one is partitioned into three categories: (1) messages which lead to trivial subforms, identifying a unique player to select an outcome in A (denote these H_0), (2) messages leading to stage 2 (denoted H_1) – to subforms where player interaction is critical, and (3) messages (the set $M \setminus (H_0 \cup H_1)$) which lead to termination of the game with outcome $x(s)$. From the description above:

$$H_1 = \{(s_k, \alpha(k), n_k)_{k \in I} | \exists i, \alpha(i) \neq \hat{\alpha}, \alpha(j) = \hat{\alpha}, j \neq i,\ s = (s_1, \ldots, s_I) = s(\alpha),\ n_k = 0, \forall k\}$$

$$H_0 = \{m \in M \mid \#\{k | n_k \geq 0\} \geq 2,\ \text{or}\ \#\{k | \alpha(k) \neq \hat{\alpha}\} \geq 2,\ \text{or both}\}$$

Furthermore, if $m = (m_1, m_2, \ldots, m_n) \in H_0$, $m_k = (s_k, \alpha(k), n_k)$, a unique i is associated to m according to $i = \min\{\arg\max_i n_i\}$. Thus, the set of messages H_0 can be partitioned into n sets, $\{H_{oi}\}_{i \in I}$ where: $H_{oi} = \{\{m_i\}_{i \in I} \in H_0 | i = \min\{\arg\max_i n_i\}\}$. H_{oi} are those messages in H_0 which lead to i being selected as dictator.

Stage 2

Suppose that Stage 2 is reached because player i announced $\alpha \neq \hat{\alpha}$ in Stage 1, $\alpha \in D(i)$ and $\bar{s} = s(\alpha)$ is the vector of types announced by the players. Let $j \in I$ and y^{α} , $z^{\alpha} \in X$ be the "other" player and the allocations respectively associated with α in the posterior reversal condition. In Stage 2, each agent must select an element in the set $B = \{R, W\} \times N \times A$. Agent i chooses a color $c_i \in \{R, W\}$ (red or white), an integer $\ell_i \in \mathcal{N}$ and an outcome $e_i \in A$.[6] Outcomes are determined as follows.

1. If $\#\{k \in I | c_k = R\} \geq n - 1$, then the outcome is a_i.
2. If $\#\{k \in I | c_k = W\} \geq n - 1$ and
   (a) $j \in \{k \in I | c_k = W\}$, then the outcome is $y^{\alpha}(\bar{s})$.
   (b) $j \notin \{k \in I | c_k = W\}$, then the outcome is $z^{\alpha}(\bar{s})$.
3. In all other cases, the outcome is e_k where $k = \min\{\arg\max_i \ell_i\}$.

The extensive form is depicted in the figure.

B. Strategies and Equilibrium

Strategies

A strategy for agent i is a triple $(f_i, g_{io}, g_i) = \sigma_i$ where $f_i : s_i \to M_i$, $g_{io} : S_i \times H_{io} \to A$ and $g_i : S_i \times H_1 \to B$. If the state chosen by the prior distribution on types is $s = (s_1, \ldots, s_n) \in S$, then agents play $f(s) = (f_1(s_1), \ldots, f_n(s_n))$ in Stage 1. If there is a double defection (some i, j pair with $\alpha(i) \neq \hat{\alpha} \neq \alpha(j)$, or $n_i, n_j > 0$ or both), the message $m \in H_0$ is in one of the sets H_{ko}, say the ith, and agent i chooses an outcome according to $a = g_{io}(s_i, m)$. If Stage 2 is reached, then given history $h \in H_1$, agents choose $g(s, h) = (g_1(s_1, h), ..., g_n(s_n, h))$, with $g_i(s_i, h) \in B$.

Sequential Equilibrium

The set of paths in the game is given by $S \times H$, where $H = (M \setminus [H_0 \cup H_1]) \cup [H_0 \cup A] \cup [H_1 \cup B]$. Each history $h \in H$ determines an outcome $a(h) \in A$. In the game, the initial history is null: $\emptyset$. Depending on the actions taken initially, the game terminates, or some history $h \in H_{oi}$ is announced and agent i asked to choose a point in A, or some history in $h \in H_1$ is announced and the game proceeds to stage 2. For agent i, the collection of information sets in the game is given by a pair (s_i, h) where $s \in S_i$ and $h \in \{\emptyset \times H_{io} \times H_1\}$.

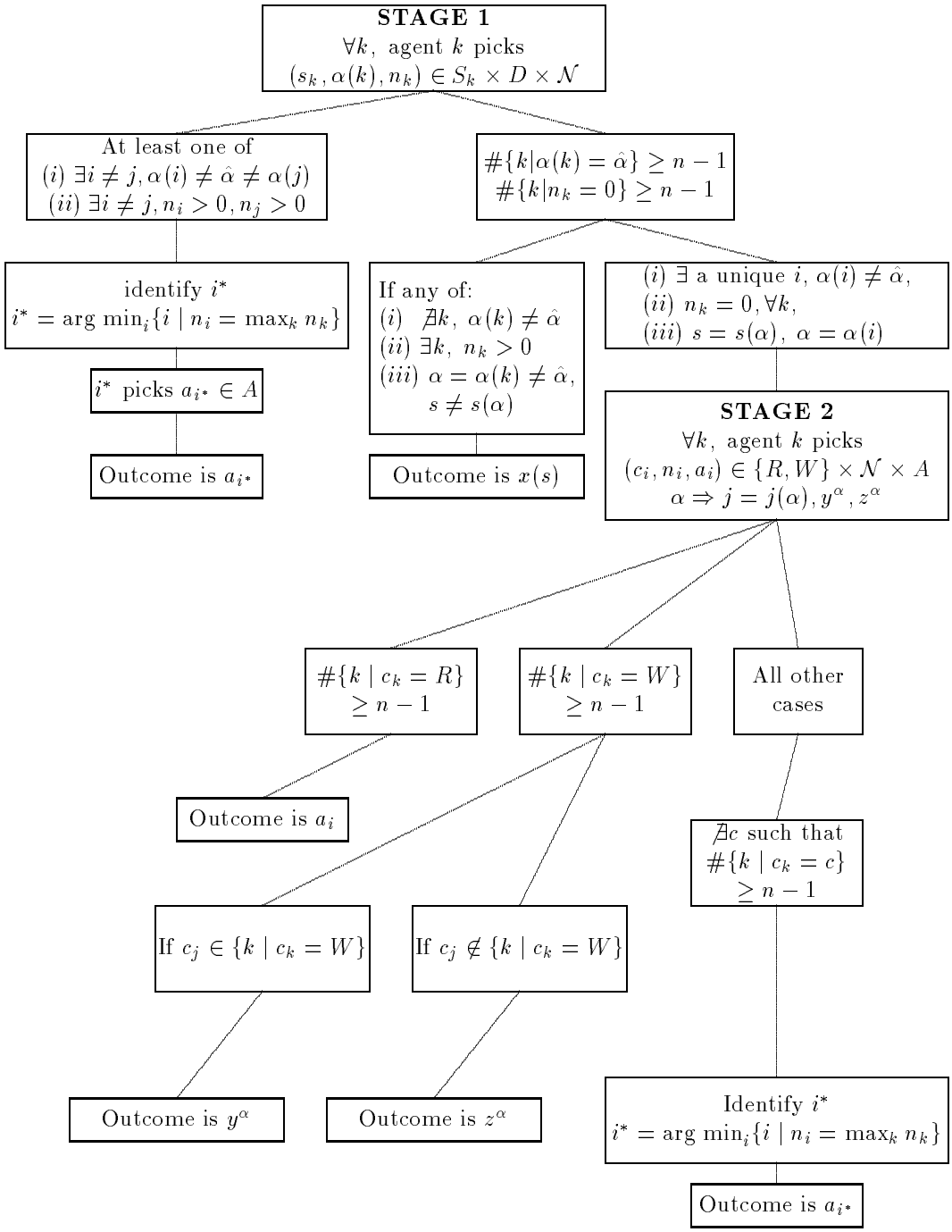
Definition 12 *A* **Sequential Equilibrium** *is a strategy σ^* and a collection of distributions $\psi = \{([\psi(s_i, h_1)]_{s_i \in S_i})_{i \in I})\}_{h_1 \in H_1}$ where $\psi(s_i, h_1)$ is the conditional distribution of agent i type s_i, on S_{-i}, conditional on h_1.*

1. *The distribution vector ψ is required to be consistent in the sense that it is obtained as the limit of a sequence of conditional distributions which are fully determined by a sequence of fully mixed strategies σ^k, which converge to σ^*.*
2. *For each i, $(h, s_i) \in H_1 \times S_i$,*

$$\mathbf{E}_{\psi, \sigma^*}\{u_i(a(h), s) \mid (h, s_i)\} \geq \mathbf{E}_{\psi, \sigma^*_{-i}, \sigma_i}\{u_i(a(h), s) \mid (h, s_i)\}, \forall \sigma_i.$$

Note that the action sets of agents are countable, so the definition of sequential equilibrium extends directly: sequential rationality is checked at every subform and beliefs are determined as limits belief sequences generated by fully mixed strategies. We now show that the game form implements the social choice function in sequential equilibrium.

[6] Using ℓ_i and e_i to distinguish second stage integers and outcomes from the first stage integers and outcomes (n_i and a_i).



C. Implementing Equilibria

We proceed in the usual way. In Step 1 we show that there is a sequential equilibrium of the game which supports x. In Step 2, we shall show that x is the unique sequential equilibrium outcome.

(1) The allocation x is an equilibrium allocation

Consider the following strategy profile (f^*, f_o^*, g^*) (and associated consistent beliefs), defined as follows:

1. $\forall\ s_i \in S_i$ and $i \in I$, $f_i^*(s_i) = (s_i, \hat{\alpha}, 0)$.
2. $\forall\ s_i \in S_i, i \in I$ and histories $h \in H_{io}$, $f_{io}(s_i, h)$ is optimal relative to the belief system.

3. $\forall s_i \in S_i, i \in I$ and histories $h \in H_1$, $g_i(s_i, h) = (W, 0, a^*)$, where a^* is some arbitrary element in A.

For all $\alpha \in D \setminus \{\hat{\alpha}\}$, let $h_1(\alpha) \in H_1$ denote the *unique* history[7] which leads to Stage 2 following player i's announcement of $\alpha \neq \hat{\alpha}$ with the type vector announcement $\bar{s} = s(\alpha)$. Let $\mu'(\alpha) \in C(s(\alpha), i, \hat{\alpha})$ be the sequentially consistent distribution chosen to satisfy part 2 of posterior reversal. Since $\bigcup_{\alpha \neq \hat{\alpha}} h_1(\alpha) = H_1$, $\{\mu'(\alpha)\}_{\alpha \neq \hat{\alpha}}$ associates a belief-system with every history $h_1 \in H_1$. We claim that $(f^*, f_o^*, g^*, \{u'(\alpha)\}_{\alpha \neq \hat{\alpha}})$ constitutes a sequential equilibrium which supports x.

The outcome of (f^*, g^*) is x. Suppose that Stage 2 has been reached following the history $h_1(\alpha)$. The strategy g^* requires all players to play W irrespective of type. This yields the allocation y^α as the outcome. The only player who can change the outcome is player j and the outcome that obtains if he deviates is z^α. But according to part 2 of posterior reversal, $y^\alpha R^j(s_j, \mu'(\alpha)) z^\alpha$ for all types of j which have positive probability under $\mu'(\alpha)$.[8] Therefore, the actions prescribed by g^* are sequentially rational relative to the beliefs $\mu'(\alpha)$. Consider a deviation where some player does not truthfully report types: $f_i(s_i) = (\alpha_i(s_i), \hat{\alpha}, n_i)$. The outcome is then $x_{(\alpha_i, \hat{\alpha}_{-i})}$. But $x R^i(s_i, \mu) x_{(\alpha_i, \hat{\alpha}_{-i})}$, according to Self-Selection.

Next consider a deviation by some player i of the kind: $f_i(s_i) = (\tilde{\alpha}_i(s_i), \alpha, n_i)$ where $\tilde{\alpha}_i \in D_i$ and $\alpha \in D(i)$. If the game goes to stage two ($\bar{s} = s(\alpha)$ is realized), then the outcome is y^α, otherwise ($\forall s \in S \setminus \{\bar{s}\}$) the outcome is $x_{(\tilde{\alpha}_i, \hat{\alpha}_{-i})}(s)$ (or $x(s')$ where $s' \neq \bar{s}$ is the announced state). Thus, the outcome over states is w^α, so the strategy of player i yields the allocation $w^\alpha_{(\tilde{\alpha}_i, \hat{\alpha}_{-i})}$. Part 1 of posterior reversal ensures that this deviation is not profitable.

Finally, note that histories in H_0 are reachable only by joint deviations of two agents. Therefore, the strategies and beliefs constitute a sequential equilibrium.

(2) The allocation x is the unique equilibrium allocation

We first claim that there can be no sequential equilibrium in which some type of player either announces a deception not equal to $\hat{\alpha}$ or a non-zero integer. Suppose that this is false. Then there is a sequential equilibrium in which player i of type s_i sends a message such that either $\alpha_i \neq \hat{\alpha}$ or $n_i \neq 0$. Let the outcome of this strategy profile be some allocation $b \in X$. Let $T = s_i \times S_{-i} \subset S$. Since the environment is economic there exists an individual $j \neq i$ and a constant allocation $c \in X$ such that $(c_T b) P^j(s_j, \mu) b$ for all $s_j \in S_j$. Note that any type s_j of player j can attain the allocation $c_T b$ be precipitating the "integer game" and announcing an integer greater than the maximum of all integers announced by all other players. This yields a history in H_0 where j is dictator (H_{oj}). The details of the argument may be found in Jackson (1991). By an analogous argument, it follows that all sequential equilibria must have the property that in Stage 2, all types of players unanimously announce either R or W.

The only candidate sequential equilibrium left to consider is the one where Stage 1 strategies are of the form: $\forall s_i \in S_i$ and $i \in I$, $f_i(s_i) = (\alpha_i(s_i), \hat{\alpha}, 0)$ for some $\alpha_i \in D_i$. In this case the game is over in Stage 1 and the outcome is x_α. Let $i \in I$ be such that $\alpha \in D(i)$, $\bar{s} = s(\alpha)$, chosen according to the posterior reversal condition. Also, let $C(\bar{s}, i, \alpha)$ be the collection of consistent distributions at the unique history $h_1 \in H_1$ determined by α. Consider the following deviation by player i : $f_i'(s_i) = (\bar{s}_i, \alpha, 0)$.[9] The game now carries over to Stage 2 with positive probability. Consider the strategy profile where all players in this subform

[7] Recall if i announces $\alpha \neq \hat{\alpha}$ stage 2 is reached only if *every* agent k announces $n_k = 0$ and $s_k = s_k(\alpha)$ and all agents other than i announce $\alpha(j) = \hat{\alpha}$.

[8] In fact, under truth-telling, if $\bar{s}$ is announced then the posterior puts probability 1 on player j type $\bar{s}_j$: only when $j = i$ is the posterior not determined on S_j.

[9] Or $f_i'(s_i) = (\alpha_i(s_i), \alpha, 0)$ since this will reach the second stage with positive probability.

choose W so that the outcome is y^α. For any system of beliefs $\mu' \in C(\bar{s}, i, \alpha)$, there exists a type of player j, t_j who is better-off defecting in order to get allocation z^α (using part 3 of posterior reversal). Therefore, the unique equilibrium in this subform is the one where all players choose R (that it is an equilibrium is immediate). As a consequence of the Stage 1 deviation by player s_i, the outcome is an allocation of the form $(w_T x_\alpha)$ where $T = \{s \in S | \alpha(s) = \bar{s}\}$ and w is a constant allocation chosen by i, type s_i. From condition 4 of posterior reversal, such an allocation exists. Player s_i will therefore deviate and the strategy profile f cannot be part of a sequential equilibrium.

6.2 Proof of Theorem 2.

Consider condition 1 in PR. For given α and the associated $i, j \in I$, $\bar{s} \in S$,

$$x\mathbf{R}^i(s_i, \mu) w_{(\alpha_i, \hat{\alpha}_{-i})}, \forall \alpha_i \in D_i,\ s_i \in S_i.$$

This can be rewritten

$$\sum_{s_{-i}} u_i(x(s_{-i}, s_i), s_{-i}, s_i)\mu(s_{-i} \mid s_i) \geq \sum_{s_{-i}} u_i(w(s_{-i}, \alpha_i(s_i)), s_{-i}, s_i)\mu(s_{-i} \mid s_i)$$

Since $w(s) = x(s), \forall s \neq \bar{s}$, the right side of this expression can be written:

$$\sum_{s_{-i}} u_i(w(s_{-i}, \alpha_i(s_i)), s_{-i}, s_i)\mu(s_{-i} \mid s_i) = \sum_{s_{-i} \neq \bar{s}_{-i}} u_i(x(s_{-i}, \alpha_i(s_i)), s_{-i}, s_i)\mu(s_{-i} \mid s_i) + u_i(w(\bar{s}_{-i}, \alpha_i(s_i)), \bar{s}_{-i}, s_i)\mu(\bar{s}_{-i} \mid s_i)$$

Adding and subtracting $u_i(x(\bar{s}_{-i}, \alpha_i(s_i)), \bar{s}_{-i}, s_i)$ from the right side of this expression gives

$$\sum_{s_{-i}} u_i(x(s_{-i}, \alpha_i(s_i)), s_{-i}, s_i)\mu(s_{-i} \mid s_i) + [u_i(w(\bar{s}_{-i}, \alpha_i(s_i)), \bar{s}_{-i}, s_i) - u_i(x(\bar{s}_{-i}, \alpha_i(s_i)), \bar{s}_{-i}, s_i)]\mu(\bar{s}_{-i} \mid s_i)$$

Thus,

$$\sum_{s_{-i}} u_i(x(s_{-i}, s_i), s_{-i}, s_i)\mu(s_{-i} \mid s_i) \geq \sum_{s_{-i}} u_i(w(s_{-i}, \alpha_i(s_i)), s_{-i}, s_i)\mu(s_{-i} \mid s_i)$$

is satisfied if and only if

$$[u_i(w(\bar{s}_i, \alpha_i(s_i)), \bar{s}_{-i}, s_i) - u_i(x(\bar{s}_{-i}, \alpha_i(s_i)), \bar{s}_{-i}, s_i)]\mu(\bar{s}_{-i} \mid s_i) \leq 0.$$

If $\alpha_i(s_i) \neq \bar{s}_i$, then

$$[u_i(w(\bar{s}_{-i}, \alpha_i(s_i)), \bar{s}_{-i}, s_i) - u_i(x(\bar{s}_{-i}, \alpha_i(s_i)), \bar{s}_{-i}, s_i)] = 0$$

and in this case the inequality is automatically satisfied. If $\alpha_i(s_i) = \bar{s}_i$, then the expression becomes

$$[u_i(y, \bar{s}_{-i}, s_i) - u_i(x(\bar{s}), \bar{s}_{-i}, s_i)]\mu(\bar{s}_{-i} \mid s_i)$$

For each $s_i \in S_i$ this must be non positive. Consequently,

$$[u_i(x(\bar{s}), \bar{s}_{-i}, s_i) - u_i(y, \bar{s}_{-i}, s_i)] \geq 0, \forall s_i \in S_i$$

Given a prior distribution with full support, condition 1 of PR is equivalent to this.

Next, consider condition 2 of PR. This condition requires that $\exists \mu' \in \mathcal{C}(\bar{s}, i, \hat{\alpha})$, such that,

$$y\mathbf{R}^j(s_j, \mu')z, \forall s_j \in supp_j \mu', \text{ when } j \neq i, \text{ and } \forall s_j \in S_j, \text{ when } j = i.$$

Given $\hat{\alpha}$, every distribution in $\mathcal{C}(\bar{s}, i, \hat{\alpha})$ has support on $\{\bar{s}_{-i}\} \times S_i$. The condition may be written

$$\sum_{s_{-j}} u_j(y, s_{-j}, s_j)\mu'(s_{-j} \mid s_j) \geq \sum_{s_{-j}} u_j(z, s_{-j}, s_j)\mu'(s_{-j} \mid s_j)$$

for some $\mu' \in \mathcal{C}(\bar{s}, i, \hat{\alpha})$. In view of the support restriction, the condition can be rewritten. For $i \neq j$, the condition becomes

$$u_j(y, \bar{s}_{-i}, s_i^*) \geq u_j(z, \bar{s}_{-i}, s_i^*), \text{ some } s_i^* \in S_i$$

For $i = j$, the condition becomes

$$u_i(y, \bar{s}_{-i}, s_i) \geq u_i(z, \bar{s}_{-i}, s_i), \ \forall s_i \in S_i$$

Finally, consider condition 3 of PR. This condition is: $\forall \mu' \in \mathcal{C}(\bar{s}, i, \alpha), \exists t_j \in supp_j \mu'$ when $j \neq i$ and $\exists t_j \in S_j$ when $j = i$ such that

$$z\mathbf{P}^j(t_j, \mu')y.$$

This can be rewritten

$$\sum_{s_{-j}} u_j(z, s_{-j}, t_j)\mu'(s_{-j} \mid t_j) > \sum_{s_{-j}} u_j(y, s_{-j}, t_j)\mu'(s_{-j} \mid t_j)$$

Given $I' \subset I$, Let $\varphi(\alpha, \bar{s}, I') = \{s_k \in S_k \mid k \notin I', \alpha_k(s_k) = \bar{s}_k\}$. This identifies the set of agents *not* in I' whose types could be mapped under α to $\bar{s}$. In particular, if $\mu' \in \mathcal{C}(\bar{s}, i, \alpha)$, then μ' has support in $\varphi(\alpha, \bar{s}, i) \times S_i$.

If $i = j$, then for some s_i, there is a $\mu' \in \mathcal{C}(\bar{s}, i, \alpha)$ with support on $\varphi(\alpha, \bar{s}, i) \times \{s_i\}$. In this case, a sufficient condition for 3 of posterior reversal to hold is that

$$\exists s_i \in S_i, \ u_i(z, s) > u_i(y, s), \ \forall s \in \varphi(\alpha, \bar{s}, i) \times \{s_i\}.$$

If $i \neq j$, then the appropriate condition becomes:

$$\exists t_j \in \alpha_j^{-1}(\bar{s}_j), u_j(z, s_{-j}, t_j) > u_j(y, s_{-j}, t_j), \ \forall s_{-j} \in \varphi(\alpha, \bar{s}, \{i, j\}) \times S_i.$$

This completes the proof of theorem 2, since condition 4 of posterior reversal is unchanged.

References

[1] Abreu, D. and H. Matsushima, (1990) "Virtual implementation in Iteratively Undominated Strategies: Incomplete Information", Mimeo, Princeton University.

[2] Abreu, D. and A. Sen, (1990) "Subgame Perfect Implementation: A Necessary And Almost Sufficient Condition", *Journal of Economic Theory*, 50.

[3] Abreu, D. and A. Sen, (1990), "Virtual Implementation in Nash Equilibrium", *Econometrica*, Vol 59.

[4] Baliga, S. (1993), "Implementation in Incomplete Information Environments: The Use of Extensive Form Games", University of Cambridge.

[5] Bergin, J and A. Sen. (1993), "Extensive Form Implementation in Incomplete Information Environments", mimeo.

[6] Bergin, J and A. Sen. (1997), "Implementing game forms with "One Round of Signaling", in incomplete information environments". mimeo.

[7] Brusco, S. (1995), "Perfect Bayesian Implementation", *Economic Theory*, Vol. 5.

[8] Brusco, S. (1997), "Perfect Bayesian Implementation: One Round of Signaling is Not Enough", mimeo, January 1997, Departamento de Economía de la Empresa, Universidad Carlos III de Madrid.

[9] Dutta, B. and A. Sen (1994), " Bayesian Implementation: The necessity of Infinite Mechanisms ", *Journal of Economic Theory*, 1994.

[10] Dutta, B. and A. Sen (1995) "Two person Bayesian Implementation", *Economic Design*, 1995.

[11] Moore, J. and R. Repullo, (1988), "Subgame Perfect Implementation", *Econometrica*, 56.

[12] Jackson, M, O, (1991), "Bayesian Implementation", *Econometrica*, vol 59, No. 2.

[13] Myerson, R. (1979), "Incentive Compatibility and the Bargaining Problem", *Econometrica*, vol 47.

[14] Palfrey, T. (1992), "Implementation in Bayesian Equilibrium: The Multiple Equilibrium Problem in Mechanism Design", in *Advances in economic theory; invited papers for the sixth world congress of the econometric society*, Vol I, ed. J.J. Laffont. Cambridge University Press.

[15] Palfrey, T. and Srivastava, S. (1989), "Mechanism Design with Incomplete Information: A solution to the Implementation Problem", *Journal of Political Economy*, vol 97.

[16] Palfrey, T. and Srivastava, S. (1989), "Nash Implementation with Incomplete Information in Exchange Economies", *Econometrica*, vol 59.

[17] Postlewaite, A. and Schmeidler, D. (1986), "Implementation in Differential Information Economies", *Journal of Economic Theory*, vol 39.

Implementing game forms with "One Round of Signaling" in incomplete information environments.

James Bergin and Arunava Sen

January 1997

We consider games with one round of signaling, briefly described in the text. Let C_i be the set of actions player i can take in stage 1, and $C = \times_{i=1}^{n} C_i$. In the first stage a strategy for i is a function $\sigma_i : S_i \to C_i$. Thus, $\sigma = (\sigma_1, \ldots, \sigma_n) : S \to C$. Although discussing pure strategies, write $\sigma(c_i; s_i)$ to the denote the probability that type s_i chooses action c_i. One action will have probability 1. We can use the notation $\sigma_i(s_i)$ to denote a pure strategy of agent i type s_i, with $\sigma_i(s_i) \in C_i$. Partition $C = C^1 \cup C^2$, where the game terminates if $c \in C^1$ is chosen, and goes to stage 2 if $c \in C^2$ is chosen. Let $g : C^1 \to A$, where A is the set of outcomes. For the discussion to follow, we focus on a two stage game, so the second stage is the final stage. However, the argument here extends almost without modification to the case of multiple stages following period 1. We explain why in a remark below. The discussion depends critically on the beliefs at the beginning of period 2, and not on the number of stages remaining in the game.

If the game goes to stage 2, then a strategy for i is a function from type and history to actions. Let B_c^i be the set of available actions for player i in stage 2, if action c is chosen in period 1. Let $B_c = \times_{i=1}^{n} B_c^i$ be the set of actions in the second stage. Thus, a strategy for i is a function $\tau_i(b_i, c, s_i)$ which gives the probability that type s_i chooses action b_i, given history c. Let $\tau_(b, c, s) = \{\tau_i(b_i, c, s_i)\}_{i=1}^{n}$. Again, with pure strategies, one action will have probability 1. We can write $\tau_i(c, s_i)$ to denote a pure strategy, at history c, type s_i, so $\tau_i(c, s_i) \in B_c^i$. Let $f(c, b) \in A$ be the outcome determined by the game at the end of stage 2 when c was chosen in the first period ($c \in C^2$) and $b \in B_c$ was chosen in period 2. To simplify notation, we will assume that when the game goes to stage 2, only the choice made in stage 2 affects the outcome. This simplifying assumption does not affect the result (see the remark below), and the notation is less bothersome. Thus, given a choice $b \in B_c$, let $f(b) \in A$ denote the outcome determined in the game in stage 2.

Proposition 2 *Let Γ be a game with one round of signaling which implements x, with implementing strategy $(\bar{\sigma}, \bar{\tau})$. Suppose that for some system of (stage two) beliefs consistent with $\bar{\sigma} \circ \alpha$ there is some $\tilde{\tau}$ yielding an equilibrium on every subform of the game (in stage 2). Then:*

a. If, under a belief system determined by $\bar{\sigma} \circ \alpha$, $\bar{\tau}$ remains an equilibrium across subforms (reachable by a deviation of just one player from $\bar{\sigma} \circ \alpha$), then x satisfies GBM at α.

b. If, under any belief system determined by $\bar{\sigma} \circ \alpha$, $\bar{\tau}$ does not define an equilibrium on each subform (reachable by a deviation of just one player from $\bar{\sigma} \circ \alpha$), then x satisfies CR.

1. The Equilibrium Strategy Yielding $x(s)$.

Let $(\bar{\sigma}, \tau)$ be an equilibrium strategy which implements the social choice rule (so that the second period is not reached.) Conditional on the second period being reached (a zero probability event), τ must determine an equilibrium on each subform of the game, relative to a conditional belief system determined by $\bar{\sigma}$.

Because $\bar{\sigma}$ is optimal in period 1 (given τ is used if period 2 is reached), the following condition must hold. For each i and for each $s_i \in S_i$, we require that for all $\hat{\sigma}_i$:

$$\sum_{s_{-i}}\sum_{c\in C^1} u_i(g(c),s)\bar{\sigma}_{-i}(c_{-i},s_{-i})\bar{\sigma}_i(c_i,s_i)\mu(s_{-i}\mid s_i)+$$
$$\sum_{s_{-i}}\sum_{c\in C^2}\{\sum_{b\in B_c} u_i(f(b),s)\tau(b,c,s)\}\bar{\sigma}_{-i}(c_{-i},s_{-i})\bar{\sigma}_i(c_i,s_i)\mu(s_{-i}\mid s_i)$$
$$\geq \tag{1}$$
$$\sum_{s_{-i}}\sum_{c\in C^1} u_i(g(c),s)\bar{\sigma}_{-i}(c_{-i},s_{-i})\hat{\sigma}_i(c_i,s_i)\mu(s_{-i}\mid s_i)+$$
$$\sum_{s_{-i}}\sum_{c\in C^2}\{\sum_{b\in B_c} u_i(f(b),s)\tau(b,c,s)\}\bar{\sigma}_{-i}(c_{-i},s_{-i})\hat{\sigma}_i(c_i,s_i)\mu(s_{-i}\mid s_i)$$

Working with pure strategies, at each s, $\sigma(c,s)$ puts probability one on some c, so we may write $\sigma(s)=c\in C$. Write $\tau(c,s)$ to denote the choice at period 2, conditional on s and choice c in period 1. Let χ_Q be the characteristic function of the event Q. Thus, $\chi_{\{(s_{-i},s_i)|\bar{\sigma}(s_{-i},s_i)\in C^1\}}$ is equal to 1 on the set of s's mapped to C^1 under $\bar{\sigma}$. The no gain from deviation condition becomes: $\forall i$, $\forall s_i$, $\forall\hat{\sigma}_i$:

$$\sum_{s_{-i}}\chi_{\{(s_{-i},s_i)|\bar{\sigma}(s_{-i},s_i)\in C^1\}}u_i(g(\bar{\sigma}(s)),s)\mu(s_{-i}\mid s_i)+$$
$$\sum_{s_{-i}}\chi_{\{(s_{-i},s_i)|\bar{\sigma}(s_{-i},s_i)\in C^2\}}u_i(f(\tau[\bar{\sigma}(s),s]),\ s)\mu(s_{-i}\mid s_i)$$
$$\geq \tag{2}$$
$$\sum_{s_{-i}}\chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(s_{-i}),\hat{\sigma}_i(s_i))\in C^1\}}u_i(g(\bar{\sigma}_{-i}(s_{-i}),\hat{\sigma}_i(s_i)),\ s)\mu(s_{-i}\mid s_i)+$$
$$\sum_{s_{-i}}\chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(s_{-i}),\hat{\sigma}_i(s_i))\in C^2\}}u_i(f(\tau[(\bar{\sigma}_{-i}(s_{-i}),\hat{\sigma}_i(s_i)),s]),\ s)\mu(s_{-i}\mid s_i)$$

Thus, the outcome under $\bar{\sigma}$ is:

$$\chi_{\{(s_{-i},s_i)|\bar{\sigma}(s_{-i},s_i)\in C^1\}}g(\bar{\sigma}(s))+\chi_{\{(s_{-i},s_i)|\bar{\sigma}(s_{-i},s_i)\in C^2\}}f(\tau[\bar{\sigma}(s),s]) \tag{3}$$

In equilibrium, with the second stage not reached $\bar{\sigma}(\cdot)$ has range C^1, $\bar{\sigma}(S)\subseteq C^1$. Thus, $\chi_{\{(s_{-i},s_i)|\bar{\sigma}(s_{-i},s_i)\in C^2\}}=0$, $\forall s\in S$, so, $x(s)=g(\bar{\sigma}(s))$, $\forall s\in S$. The deviation $\hat{\sigma}_i$ gives $g((\bar{\sigma}_{-i},\hat{\sigma}_i)(s))$ if the game terminates at stage 1 and $f(\tau[(\bar{\sigma}_{-i},\hat{\sigma}_i)(s),s])$, if the game goes to stage 2. So, the outcome resulting from the deviation is:

$$\chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(s_{-i}),\hat{\sigma}_i(s_i))\in C^1\}}g(\bar{\sigma}_{-i}(s_{-i}),\hat{\sigma}_i(s_i))+$$
$$\chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(s_{-i}),\hat{\sigma}_i(s_i))\in C^2\}}f(\tau[(\bar{\sigma}_{-i}(s_{-i}),\hat{\sigma}_i(s_i)),s]) \tag{4}$$

2. Eliminating Deceptions as Equilibria.

Next, we consider an alternative strategy in the game of the form $\tilde{\sigma}(s)=\bar{\sigma}(\alpha(s))$, where α is a deception. Because $\bar{\sigma}(S)\subseteq C^1$, $\bar{\sigma}(\alpha(S))\subseteq C^1$. Thus, the outcome determined by $\tilde{\sigma}$ at s is $g(\bar{\sigma}(\alpha(s)))$, and since $g(\bar{\sigma}(s))=x(s)$, $g(\bar{\sigma}(\alpha(s)))=x(\alpha(s))$. Thus, unless $x(s)=x(\alpha(s)),\forall s\in S$, $\tilde{\sigma}$ cannot be the first period component of any equilibrium strategy $(\tilde{\sigma},\tilde{\tau})$. There are two possibilities.

A. The first is that there is some consistent belief system determined from $\tilde{\sigma}$ such that $\exists\tilde{\tau}$ where on each subform, $\tilde{\tau}$ defines an equilibrium on that subform relative to the consistent beliefs.

B. The second is that no such $\tilde{\tau}$ exists: for any collection of beliefs across subforms (determined by $\tilde{\sigma}$), there is some subform with no equilibrium relative to the associated beliefs. If the second possibility holds, then for every τ', $(\tilde{\sigma}, \tau')$ is not an equilibrium and the deception is "knocked out" as an equilibrium.

We begin with a discussion of case **A**, and consider **B** at the end.

A. Suppose that $(\tilde{\sigma}, \tilde{\tau})$ is an equilibrium. This implies that there is an improving deviation for some type of some player that must occur in the first stage. For, if no deviation occurs in stage 1, then the outcome is $g \circ \alpha$ (and the second stage is unreached by $\tilde{\sigma}$.) Thus, given $(\tilde{\sigma}, \tilde{\tau})$, where $\tilde{\tau}$ is an equilibrium strategy determining behavior from stage two on, some i type, s_i, wishes to deviate. Concerning $\tilde{\tau}$ there are two cases to consider:

A1. $\tilde{\tau}$ agrees with τ at each subform (or at each subform reached under the relevant challenge – we clarify below), and

A2. $\tilde{\tau}$ differs from τ on some subform (or at some subform reached under the relevant challenge – again the discussion below will clarify). We start with case **A1**.

<u>Case **A1.**</u> Suppose $\tilde{\tau} = \tau$. So, $\exists i,\ s_i \in S_i,\ \hat{\sigma}_i$

$$\begin{aligned}
&\sum_{s_{-i}} \sum_{c \in C^1} u_i(g(c), s) \tilde{\sigma}_{-i}(c_{-i}, s_{-i}) \tilde{\sigma}_i(c_i, s_i) \mu(s_{-i} \mid s_i) + \\
&\qquad \sum_{s_{-i}} \sum_{c \in C^2} \{ \sum_{b \in B_c} u_i(f(b), s) \tau(b, c, s) \} \tilde{\sigma}_{-i}(c_{-i}, s_{-i}) \tilde{\sigma}_i(c_i, s_i) \mu(s_{-i} \mid s_i) \\
&< \\
&\sum_{s_{-i}} \sum_{c \in C^1} u_i(g(c), s) \tilde{\sigma}_{-i}(c_{-i}, s_{-i}) \hat{\sigma}_i(c_i, s_i) \mu(s_{-i} \mid s_i) + \\
&\qquad \sum_{s_{-i}} \sum_{c \in C^2} \{ \sum_{b \in B_c} u_i(f(b), s) \tau(b, c, s) \} \tilde{\sigma}_{-i}(c_{-i}, s_{-i}) \hat{\sigma}_i(c_i, s_i) \mu(s_{-i} \mid s_i)
\end{aligned} \tag{5}$$

Rearranging slightly,

$$\begin{aligned}
&\sum_{s_{-i}} \chi_{\{(s_{-i}, s_i) \mid (\tilde{\sigma}_{-i}(s_{-i}), \tilde{\sigma}_i(s_i)) \in C^1\}} u_i(g(\tilde{\sigma}_{-i}(s_{-i}), \tilde{\sigma}_i(s_i)), s) \mu(s_{-i} \mid s_i) + \\
&\qquad \sum_{s_{-i}} \chi_{\{(s_{-i}, s_i) \mid (\tilde{\sigma}_{-i}(s_{-i}), \tilde{\sigma}_i(s_i)) \in C^2\}} u_i(f(\tau[(\tilde{\sigma}_{-i}(s_{-i}), \tilde{\sigma}_i(s_i)), s]), s) \mu(s_{-i} \mid s_i) \\
&< \\
&\sum_{s_{-i}} \chi_{\{(s_{-i}, s_i) \mid (\tilde{\sigma}_{-i}(s_{-i}), \hat{\sigma}_i(s_i)) \in C^1\}} u_i(g(\tilde{\sigma}_{-i}(s_{-i}), \hat{\sigma}_i(s_i)), s) \mu(s_{-i} \mid s_i) + \\
&\qquad \sum_{s_{-i}} \chi_{\{(s_{-i}, s_i) \mid (\tilde{\sigma}_{-i}(s_{-i}), \hat{\sigma}_i(s_i)) \in C^2\}} u_i(f(\tau[(\tilde{\sigma}_{-i}(s_{-i}), \hat{\sigma}_i(s_i)), s]), s) \mu(s_{-i} \mid s_i)
\end{aligned} \tag{6}$$

The outcome with the deception, (and i playing $\tilde{\sigma}_i$) is

$$\begin{aligned}
&\chi_{\{(s_{-i}, s_i) \mid (\tilde{\sigma}_{-i}(s_{-i}), \tilde{\sigma}_i(s_i)) \in C^1\}} g(\tilde{\sigma}_{-i}(s_{-i}), \tilde{\sigma}_i(s_i)) + \\
&\qquad \chi_{\{(s_{-i}, s_i) \mid (\tilde{\sigma}_{-i}(s_{-i}), \tilde{\sigma}_i(s_i)) \in C^2\}} f(\tau[(\tilde{\sigma}_{-i}(s_{-i}), \tilde{\sigma}_i(s_i)), s])
\end{aligned} \tag{7}$$

Since $\tilde{\sigma}(S) = \bar{\sigma}(\alpha(S)) \subseteq \bar{\sigma}(S) \subseteq C^1$, the second stage is not reached. Thus, the outcome is $g(\tilde{\sigma}_{-i}(s_{-i}), \tilde{\sigma}_i(s_i)) = g(\tilde{\sigma}(s)) = g(\bar{\sigma}(\alpha(s)))$. Now, since there is some player who wishes to deviate, suppose that type s_i of player i deviates to upset the deception and plays $\hat{c}_i = \hat{\sigma}_i(s_i)$.

Refer to expression 4, and note that if player i, type s_i plays $\hat{c}_i = \hat{\sigma}_i(s_i)$, then expression 4 may be written:

$$\chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(s_{-i}),\hat{c}_i)\in C^1\}} g(\bar{\sigma}_{-i}(s_{-i}),\hat{c}_i) + \chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(s_{-i}),\hat{c}_i)\in C^2\}} f(\tau[(\bar{\sigma}_{-i}(s_{-i}),\hat{c}_i),s]) \tag{8}$$

With the deception, the "challenge" $\hat{c}_i$ produces the outcome:

$$\chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(\alpha(s_{-i})),\hat{c}_i)\in C^1\}} g(\bar{\sigma}_{-i}(\alpha(s_{-i})),\hat{c}_i) + \chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(\alpha(s_{-i})),\hat{c}_i)\in C^2\}} f(\tau[(\bar{\sigma}_{-i}(\alpha(s_{-i})),\hat{c}_i),s]) \tag{9}$$

Define the function:

$$y(s,s) \equiv \chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(s_{-i}),\hat{c}_i)\in C^1\}} g(\bar{\sigma}_{-i}(s_{-i}),\hat{c}_i) + \chi_{\{(s_{-i},s_i)|(\bar{\sigma}_{-i}(s_{-i}),\hat{c}_i)\in C^2\}} f(\tau[(\bar{\sigma}_{-i}(s_{-i}),\hat{c}_i),s]) \tag{10}$$

Or, to clearly identify the role of each coordinate:

$$y(t,s) \equiv \chi_{\{(t_{-i},t_i)|(\bar{\sigma}_{-i}(t_{-i}),\hat{c}_i)\in C^1\}} g(\bar{\sigma}_{-i}(t_{-i}),\hat{c}_i) + \chi_{\{(t_{-i},t_i)|(\bar{\sigma}_{-i}(t_{-i}),\hat{c}_i)\in C^2\}} f(\tau[(\bar{\sigma}_{-i}(t_{-i}),\hat{c}_i),s]) \tag{11}$$

Note that $y(s,s)$ is the outcome at state s produced by a challenge of player i type s_i in the equilibrium achieving x (see equations 4, 8 and 9). Note also that $y(s,s)$ is independent of s_i in the first coordinate position (or t_i in the clarifying notation). Also, observe that $y(\alpha(s),s)$ is the outcome produced by a challenge from player i type s_i in the deception (see equation 9.) Write y_α to denote the function $y(\alpha(s),s)$, $s \in S$. Similarly, write y_{α_i} to denote the function $y((\alpha_i(s_i),s_{-i}),s)$. In view of the definition of y, $y = y_{\alpha_i(s_i)}$ (again because $y(s,s)$ is independent of s_i in the first coordinate position).

Since, in the equilibrium, x is preferable to player i type s_i than what is achieved through deviation, y or y_{α_i} or $y_{\alpha_i(s_i)}$ (since $y = y_{\alpha_i} = y_{\alpha_i(s_i)}$), we have that for any type of player i:

$$x R^i(s_i,\mu) y_{\alpha_i}(s_i) \tag{12}$$

However, in the deception, player i type s_i wishes to deviate:

$$y_\alpha P^i(s_i,\mu) x_\alpha \tag{13}$$

Call this condition generalized Bayesian monotonicity. Formally:

Definition 13 *A social choice rule, x, satisfies* **generalized Bayesian monotonicity** *(GBM) if given $\alpha \in D$, $x_\alpha \neq x$, $\exists i \in I, s_i \in S_i$ and $y : S \times S \to A$ such that:*

a. $x R^i(t_i,\mu) y_{\alpha_i(s_i)}$, $\forall t_i \in S_i$

b. $y_\alpha P^i(s_i,\mu) x_\alpha$

We now turn to the second case, **A2**.

<u>Case **A2.**</u> Suppose $\tilde{\tau} \neq \tau$. The previous calculations were based on τ being an equilibrium relative to beliefs determined by $\tilde{\sigma}$. However, the strategy $\bar{\sigma}(\alpha(\cdot))$ will generate different posterior distributions in the second stage compared to those determined by $\bar{\sigma}(\cdot)$, and these may not admit the strategies induced by τ on all subforms as equilibria on those subforms. Suppose that an equilibrium on each subform exists (relative to beliefs determined by $\tilde{\sigma}$) and is given by some function $\tilde{\tau} \neq \tau$. Again, a deviation must occur in stage 1 (if no

one in stage 1 wishes to deviate from $\tilde{\sigma}$ and $\tilde{\tau}$ gives an equilibrium on each subform for beliefs determined by $\tilde{\sigma}$, then $(\tilde{\sigma}, \tilde{\tau})$ is an equilibrium with outcome $g \circ \alpha$.) So, for some type, s_i, of some player i, there is a profitable deviation, $\hat{c}_i = \hat{\sigma}_i(s_i)$, from $\tilde{\sigma}_i(s_i)$ in stage 1. If $\tilde{\tau}$ and τ agree on each $c \in C^2 \cap \{c \mid \exists t_{-i}, (\tilde{\sigma}_{-i}(t_{-i}), \hat{c}_i) = c\}$, (so that τ is an equilibrium on each subform reached with positive probability under the deviation), then the calculations are exactly as before and we obtain conditions (12) and (13). So, suppose that for the profitable deviation $\hat{c}_i$ (by player i type s_i), there is some $s_{-i} \in S_{-i}$ such that at the subform reached by $c^* = (\tilde{\sigma}_{-i}(s_{-i}), \hat{c}_i)$, τ and $\tilde{\tau}$ are necessarily different: the beliefs on this subform do not admit τ as an equilibrium. This means that, at this subform, given consistent beliefs determined by $\tilde{\sigma}$ some type s_j of some player j has a profitable deviation on that subform, relative to the strategy determined there by τ (and no profitable deviation from τ at any subform under the beliefs determined by $\bar{\sigma}$ that support τ). Let τ_j^* be such a second stage strategy for j and put $\tau^* = (\tau_{-j}, \tau_j^*)$. Define

$$z(t,s) \equiv \chi_{\{(t_{-i},t_i)|(\bar{\sigma}_{-i}(t_{-i}),\hat{c}_i)\in C^1\}} g(\bar{\sigma}_{-i}(t_{-i}), \hat{c}_i) + \chi_{\{(t_{-i},t_i)|(\bar{\sigma}_{-i}(t_{-i}),\hat{c}_i)\in C^2\}} f((\tau^*[(\bar{\sigma}_{-i}(t_{-i}), \hat{c}_i), s]) \tag{14}$$

Thus, $z(s,s)$ is the outcome at type profile s with truthful reporting in stage 1, a challenge $\hat{c}_i$ by player i type s_i and a deviation by player j from τ_j to τ_j^* in stage 2. Note that z differs from y only in that j chooses τ_j^* rather than τ_j in stage 2. By assumption, $y_{\alpha_i(s_i)} R^j(s_j, \mu) z_{\alpha_i(s_i)}$: under $\bar{\sigma}$ on each stage 2 subform, τ defines an equilibrium — no type of any player (including j) has an incentive to deviate. But, since the subform associated with c^* (where j, type t_j's preference flip occurs) has positive probability under $(\tilde{\sigma}_{-i}, \hat{c}_i)$, $z_\alpha P^j(t_j, \mu) y_\alpha$. Thus, in this case, the following condition holds:

Definition 14 *A social choice rule, x, satisfies* **chain reversal** *(CR) if given $\alpha \in D$, $x_\alpha \neq x$, $\exists i, j \in I$, $s_i \in S_i$, $s_j \in S_j$ and functions y and z, $y : S \times S \to A$, $z : S \times S \to A$,*

a. $xR^i(t_i, \mu) y_{\alpha_i(s_i)}$, $\forall t_i \in S_i$,

b. $y_{\alpha_i(s_i)} R^j(t_j, \mu) z_{\alpha_i(s_i)}$, $\forall t_j \in S_j$

c. $z_\alpha P^j(s_j, \mu) y_\alpha$

Finally, note that since there is an equilibrium on each subform under some system of beliefs determined by $\bar{\sigma} \circ \alpha$. Let $\tilde{\tau}$ be the stage two equilibrium strategy. Define (noting the presence of $\tilde{\tau}$):

$$w(t,s) \equiv \chi_{\{(t_{-i},t_i)|(\bar{\sigma}_{-i}(t_{-i}),\hat{c}_i)\in C^1\}} g(\bar{\sigma}_{-i}(t_{-i}), \hat{c}_i) + \chi_{\{(t_{-i},t_i)|(\bar{\sigma}_{-i}(t_{-i}),\hat{c}_i)\in C^2\}} f(\tilde{\tau}[(\bar{\sigma}_{-i}(t_{-i}), \hat{c}_i), s]) \tag{15}$$

Given that $\tilde{\tau}$ is the strategy used on the subforms, a deception in conjunction with the challenge $\hat{c}_i$ produces the outcome $w(\alpha(s), s)$ or w_α. Assuming Bayesian monotonicity fails, the deception is upset with a challenge that causes play to reach stage 2 with positive probability. For player i (type s_i) to challenge in this way requires:

$$w_\alpha P^i(s_i, \mu) x_\alpha \tag{16}$$

This completes the proof. ■

Remark 10 Here, it is easy to see that the specification of the subforms to be independent of the first stage choice is irrelevant. If the function f depended on the first stage choice, then for example, in equation 8 we would write

$$f(\tau[(\bar{\sigma}_{-i}(s_{-i}), \hat{c}_i), s], (\bar{\sigma}_{-i}(s_{-i}), \hat{c}_i))$$

so dependence on the history is both direct and indirect (through τ). With this modification, define y as before, and the calculations are unchanged.

Remark 11 Note that the number of stages in the game is irrelevant. The issues revolve around the payoffs a deviator receives conditional on reaching a second stage – regardless of the number of subsequent stages there might be in the game.