Subgame-Perfect Implementation Under Value Perturbations*

Philippe Aghion, Drew Fudenberg, Richard Holden,
Takashi Kunimoto and Olivier Tercieux†

October 15, 2010

Abstract

We consider the robustness of extensive form mechanisms when common knowledge of the state of Nature is relaxed to common p-beliefs about it. We show that with even an arbitrarily small amount of such uncertainty, the Moore-Repullo mechanism does not yield (even approximately) truthful revelation and in addition there are sequential equilibria with undesirable outcomes. More generally, we show that any extensive form mechanism is fragile in the sense that if a non-monotonic social objective can be implemented with this mechanism, then there are arbitrarily small common p-belief value perturbations under which an undesirable sequential equilibrium exists.

1 Introduction

The literature on complete-information implementation supposes that players know the payoff-relevant state of the world, and asks which mappings from states to outcomes, i.e which social choice rules, can be implemented by mechanisms that respect the players' incentives. Although only monotonic social rules are "Nash implementable" (Maskin, 1999), a larger class of social choice rules can be implemented in extensive form games provided that a more restrictive equilibrium notion is used.

*This paper builds on two preliminary contributions, respectively by Aghion, Fudenberg and Holden (2009) and Kunimoto and Tercieux (2009).

†Aghion: Harvard University, Department of Economics; email: paghion@fas.harvard.edu. Fudenberg: Harvard University, Department of Economics; email: dfudenberg@harvard.edu. Holden: University of Chicago, Booth School of Business; email: richard.holden@chicagobooth.edu. Kunimoto: Department of Economics, McGill University and CIREQ, Montreal, Quebec, Canada; email takashi.kunimoto@mcgill.ca. Tercieux: Paris School of Economics, Paris, France; email: tercieux@pse.ens.fr. We thank Oliver Hart, Johannes Horner, John Moore and Andy Skrzypacz for detailed comments on earlier drafts. We are also grateful to Ken Binmore, Yeon-Koo Che, Mathias Dewatripont, Bob Gibbons, Ed Green, Matt Jackson, Philippe Jehiel, Hitoshi Matsushima, Eiichi Miyagawa, Roger Myerson, Andrew Postlewaite, Jean Tirole, Jorgen Weibull, Ivan Werning, Tom Wilkening and Muhamet Yildiz and seminar participants at Chicago Booth, Harvard, the Paris School of Economics, Stockholm University, the Stockholm School of Economics, Simon Fraser University, Boston University, Bocconi University, the Max Planck Institute in Bonn, and the Canadian Institute for Advanced Research, for very useful comments and suggestions.

This paper considers the robustness of subgame perfect implementation by extensive form mechanisms to common p-belief value perturbations, which are situations where each player believes with probability p slightly less than 1 that other players believe with probability slightly less than 1...and so on, that the state of nature is equal to a particular θ.[1] It is known that refinements of Nash equilibrium are not robust to allowing for general (small) perturbations of the information structure (e.g, see Fudenberg, Kreps and Levine (1988), henceforth FKL). The significance of "value perturbations" is that we fix the map from states to payoffs and only perturb the agents' beliefs over the fixed space Θ of states of the world, so that the set of messages in the mechanism remain cheap talk and do not enter directly into the payoff functions.

Our starting point is the Moore and Repullo (1988) (MR) result which roughly says that for any social choice function, one can design a mechanism that yields unique implementation in subgame perfect equilibria (i.e. for all states of nature, the set of all subgame perfect equilibria of the induced game yields the desired outcome). In particular, in environments with money, Moore and Repullo propose a simple mechanism inducing truth-telling as the unique subgame perfect equilibrium. As in MR, our focus is on *full* and *exact* implementation: full implementation means that our search for mechanisms whose entire set of equilibrium outcomes relates to the given social choice rule; exact implementation refers to the fact that we require the set of equilibrium outcomes to *exactly* coincide with those picked by the rule. Let us thus stress from the outset that exact implementation will be the notion of implementation sought in the current paper.[2]

The requirement of exact implementation can be decomposed into the following two parts: (1) there always exists an equilibrium whose outcome coincides with the given rule; (2) there are no undesirable equilibria in the sense that the undesirable equilibrium outcomes do not coincide with the given rule.

Our first result is concerned with the non-robustness of the first requirement of exact implementation: namely, whenever a MR mechanism implements a non-monotonic SCF, the truth-telling equilibrium ceases to be an equilibrium in some nearby environment. More specifically, it says that a Moore-Repullo mechanism which implements a social choice function (SCF) f under common knowledge does not yield even approximately truthful revelation in common p-belief value perturbations of the information structure if this SCF is not Maskin-monotonic.[3] In addition, we show

[1] Monderer and Samet (1989) introduced the idea of common p-belief perturbations over a general state space; here we apply their idea of beliefs on the distribution of θ. This is a "smaller" perturbation and less demanding than the one used for instance in Oury and Tercieux (2009). See also Kunimoto (2010) for a characterization of the perturbation used in this paper.

[2] Exact implementation is the approach taken by much of the implementation literature. An alternative approach, referred to as "virtual implementation" (see Matsushima (1988) and Abreu and Sen (1991)), uses non-deterministic mechanisms and only requires the SCF to be implemented with high probability. As pointed out by Jackson (2001), unlike exact implementation, virtual implementation is not robust to introducing a small amount of nonlinearity in preferences over lotteries. Moreover, as pointed out by Jackson (2001) , virtual implementation typically induces renegotiation to occur on the equilibrium path. In fact Abreu and Matsuhima (1992) already acknowledge that, in the context of virtual implementation, the mechanism designer has to commit ex ante to choosing outcomes that may turn out to be arbitrarily inefficient, even though the probability of such an event is small.

[3] Recall that a SCF f is Maskin-monotonic if for any pair of states θ and θ' such that $f(\theta)$ never goes down in the preference ranking of any agent when moving from state θ to state θ', then necessarily $f(\theta) = f(\theta')$. As we shall stress in Section 2.5 below, this is precisely the property that SCFs usually considered in contract theory do not

that under common p-belief perturbation, there will always exist a sequential equilibrium yielding undesirable outcomes.

We then move beyond MR mechanisms to consider *any* extensive-form mechanism. Our second result is concerned with the non-robustness of the second requirement of exact implementation: namely, whenever "any" mechanism implements a non-monotonic SCF, there exists an undesirable equilibrium in some nearby environment. More specifically, restricting attention to environments with a finite state space, and to mechanisms with finite strategy[4] spaces, then given any mechanism that "subgame-perfect" implements a non-monotonic SCF f under common knowledge (i.e. whose subgame perfect equilibrium outcomes in any state θ is precisely equal to $f(\theta)$), we can find a sequence of common p-belief value perturbations of this mechanism and a corresponding sequence of sequential equilibria whose outcomes do not converge to $f(\theta)$ for at least one state θ. In other words, there always exists arbitrarily small common p-belief value perturbations under which an "undesirable" sequential (and hence perfect Bayesian) equilibrium exists.

Two basic insights underlie our analysis. The first is that even a small amount of uncertainty about the state at the interim stage, when players have observed their signals but not yet played the game, can loom large ex post once the extensive form game has started and players can partly reveal their private signals through their strategy choice at each node of the game. The second insight is that perturbations of beliefs about the underlying values can turn the outcome of a non-sequential Nash equilibrium of the game with common knowledge of θ into the outcome of a sequential equilibrium of the perturbed game. In particular, we know that any extensive-form mechanism that "subgame-perfect" implements a non-monotonic SCF under common knowledge has at least one Nash equilibrium which is not a sequential equilibrium; we prove that this undesirable Nash equilibrium can be turned into an undesirable sequential equilibrium by introducing common p-belief value perturbations.

This latter insight departs in an important way from the literature on the robustness of refinements of Nash equilibrium to (small) perturbations of the information structure. Thus for example FKL allow the perturbed games to have any possible map from terminal nodes to payoffs, so that for example some (low-probability) types might be "crazy types" with a systematic preference for truthtelling or for another particular strategy. In mechanism design, however, messages and outcome functions are not primitives but rather endogenous objects to be chosen by the social planner, so it may seem natural to restrict the perturbations to be independent of the set of messages. Our analysis is even more restrictive, and only allows perturbations to beliefs about θ; this restriction only strengthens our points about the non-robustness of extensive-form implementation.

Our paper contributes most directly to the mechanism design literature, starting with Maskin (1999)'s Nash implementation result and Moore-Repullo (1988)'s subgame perfect implementation analysis, by showing the non-robustness of subgame perfect implementation to value perturbations.[5]

satisfy.

[4] The Appendix relaxes this to countable message spaces where best responses are well defined.

[5] Other related mechanism design papers include Cremer and McLean (1988), Johnson, Pratt and Zeckhauser (1990), and Fudenberg, Levine and Maskin (1991). These papers show how one can take advantage of the correlation

Our paper is also related to Chung and Ely (2003)'s study of the robustness of undominated Nash implementation. Chung and Ely show that if a social choice function is non-monotonic but can be implemented in undominated Nash equilibrium under complete information, then there are common p-belief value perturbations under which an undesirable undominated Nash equilibrium appears. In contrast, we consider extensive-form mechanisms and show that only monotonic social choice rules can be implemented in the closure of the sequential equilibrium correspondence. In general, the existence of a bad sequential equilibrium in the perturbed game neither implies nor is implied by the existence of a bad undominated Bayesian Nash equilibrium, as undominated Nash equilibria need not be sequential equilibria, and sequential equilibria can use dominated strategies.[6]

Our paper also relates to the literature on the hold-up problem. Grossman and Hart (1986) argue that in contracting situations where states of nature are observable but not verifiable, asset ownership (or vertical integration) could help limit the extent to which one party can be held up by the other party, which in turn should encourage ex ante investment by the former. However, vertical integration as a solution to the hold-up problem has been questioned in papers which use or extend the subgame-perfect implementation approach of Moore and Repullo (1988).[7] In particular, Maskin and Tirole (1999a), henceforth MT, show that the non-verifiability of states of nature can be overcome using a 3-stage subgame perfect implementation mechanism which induces truth-telling by all parties as the unique equilibrium outcome, and does so in pure strategies. We contribute to that debate by showing that the introduction of common p-belief value perturbations is quite effective in reducing the power of subgame perfect implementation.

The paper is organized as follows. Section 2 uses a simple buyer-seller example to introduce the MR mechanism, to show why truthful implementation using this mechanism is not robust to small common p-belief value perturbations, and also why such perturbations generate an undesirable sequential equilibrium. Section 3 extends our analysis to general Moore-Repullo mechanisms with n states of nature and transferable utility, and shows for a given social choice function, truth-telling equilibria are only robust to small perturbations of the information structure if the social choice function is strategy proof (which in turn implies Maskin monotonicity under weak assumptions on preferences). In Section 4 we ask whether *any* extensive form mechanism is robust to the common p-belief value perturbations. There we prove that for any non Maskin-monotonic social choice function, one can find small common p-belief value perturbations under which an undesirable sequential equilibrium exists. Finally Section 5 concludes with a few remarks and also suggestions for future research.

between agents' signals in designing incentives to approximate the Nash equilibrium under perfect information. These papers consider static implementation games with commitment, and look at fairly general information structures, as opposed to our focus on the robustness of subgame-perfect implementation to small perturbations from complete information.

[6] Trembling-hand perfect equilibria cannot use dominated strategies, and sequential and trembling-hand perfect equilibria coincide for generic assignments of payoffs to terminal nodes (Kreps and Wilson [1982]), but the generic payoffs restriction rules out our assumption that messages are cheap talk.

[7] For example, see Aghion-Dewatripont-Rey (1994) and Maskin-Tirole (1999a, 1999b).

2 A Hart-Moore (HM) example of the Moore-Repullo mechanism

2.1 Basic setup

Consider the following simple example from Hart and Moore (2003). This example captures, in the simplest possible setting, the logic of Moore and Repullo (1988)'s subgame perfect implementation mechanism.

There are two parties, a B(uyer) and a S(eller) of a single unit of an indivisible good. If trade occurs then B's payoff is

$$V_B = \theta - p,$$

where p is the price. S's payoff is

$$V_S = p,$$

thus we normalize the cost of producing the good to zero.

The good can be of either high or low quality. If it is high quality then B values it at 14, and if it is low quality then 10. We seek to implement the social choice function whereby the good is always traded ex post, and where the buyer always pays the true θ to the seller.

2.2 Common knowledge

Suppose first that the quality θ is observable and common knowledge to both parties. Even though θ is not verifiable by a court, and therefore no initial contract between the two parties can be made credibly contingent upon θ, yet truthful revelation of θ by the buyer B and the implementation of the above social choice function, can be achieved through the following Moore-Repullo (MR) mechanism:

1. B announces either "high" or "low". If "high" then B pays S a price equal to 14 and the game then stops.
2. If B announces "low" and S does not "challenge" B's announcement, then B pays a price equal to 10 and the game stops.
3. If S challenges B's announcement then:
   (a) B pays a fine F to T (a third party)
   (b) B is offered the good for 6
   (c) If B accepts the good then S receives F from T (and also the 6 from B) and we stop.
   (d) If B rejects at 3b then S pays F to T
   (e) B and S Nash bargain 50:50 over the good.

When the true value of the good is common knowledge between B and S this mechanism yields truth-telling as the unique subgame perfect (and also sequential) equilibrium. To see this, let the

true valuation be 14, and let $F = 9$. If B announces "high" then B pays 14 and we stop. If, however, B announces "low" then S will challenge because at stage 3a B pays 9 to T and, this cost being sunk, B will still accept the good for 6 at stage 3b (since it is worth 14). Anticipating this, S knows that if she challenges B she receives $9 + 6 = 15$, which is greater than the 10 that she would receive if she did not challenge. Moving back to stage 1, if B lies and announces θ'' when the true state is θ', he gets $14 - 9 - 6 = -1$, whereas he gets $14 - 14 = 0$ if he tells the truth. It is straightforward to verify that truthtelling is also the unique equilibrium if $\theta = 10$. In that case S will not challenge B when B (truthfully) announces θ'', because now B will refuse the good at price 6 (accepting the good at 6 would yield surplus $10 - 6 = 4$ to B whereas by refusing the good and relying on Nash bargaining instead B can secure a surplus equal to $10/2 = 5$). Anticipating this, S will not challenge B because doing so would give her a net surplus equal to $10/2 - 9 = -4$ which is less than the 10 she receives if she does not challenge B's announcement.

Hence, the mechanism described here (and more generally, the Moore-Repullo mechanisms we will describe in Section 3) has two nice and important properties. First, it yields unique implementation in subgame perfect equilibrium, i.e. for any state of nature, there is a unique subgame perfect equilibrium which yields the right outcome. Second, in each state, the unique subgame perfect equilibrium is appealing from a behavioral point of view since it involves telling the truth. We will show in what follows that both of these properties fail once we introduce small perturbations to the players' beliefs about θ.

2.3 The failure of truth-telling with perturbed beliefs about value

2.3.1 Pure strategy equilibria

Note first that from Fudenberg and Tirole (1991) we know that in multi-period games where each player has only two possible types, the sets of perfect Bayesian equilibria and sequential equilibria coincide. Thus here, w.l.o.g we can focus attention on perfect Bayesian equilibria (PBE), broadly defined as a pair (μ, m) where μ is a belief profile and m is a strategy profile such that: (i) Bayes' rule is used to update beliefs whenever possible; (ii) the strategies are sequentially rational, which in this setting implies that for each period t and history h^{t-1} up to $t-1$, strategies m^t are a Bayesian equilibrium for the continuation game given beliefs $\mu(\cdot|h^{t-1})$.[8] Now, as above, suppose that the good has possible values, 14 and 10. Now suppose that the players have a common prior μ, with $\mu(\theta') = 1 - \alpha$ and $\mu(\theta'') = \alpha$ where $\alpha \in (0, 1)$, and that each player receives an independent draw from a signal structure with two possible signals s'or s'', where s' is a high signal highly correlated with θ', and s'' is a low signal highly correlated with θ''. We use the notation s'_B (resp. s''_B) to refer to the event in which B receives the high signal s' (resp. the low signal s''). The following table shows the joint probability distribution ν^ε over θ, the buyer's signal s_B, and the seller's signal s_S:

[8] A sequential equilibrium is a pair (μ, m) where μ is a belief profile and m is a strategy profile such that: (i) there is a sequence of totally mixed strategy profiles $m^n \to m$ such that the beliefs μ^n computed from m^n using Bayes' rule converge to μ; (ii) the strategies are sequentially rational, i.e for each period t and history h^{t-1} up to $t-1$, the continuation strategies are a Bayesian equilibrium for the continuation game given the beliefs $\mu(\cdot|h^{t-1})$.

ν^ε	s'_B, s'_S	s'_B, s''_S	s''_B, s'_S	s''_B, s''_S
θ'	$(1-\alpha)\left(1-\varepsilon-\varepsilon^2\right)$	$(1-\alpha)\varepsilon$	$(1-\alpha)\varepsilon^2/2$	$(1-\alpha)\varepsilon^2/2$
θ''	$\alpha\varepsilon^2/2$	$\alpha\varepsilon^2/2$	$\alpha\varepsilon$	$\alpha\left(1-\varepsilon-\varepsilon^2\right)$

Note that as ε converges to zero, $\nu^\varepsilon \to \mu$. Note also that under this signal structure the buyer's signal becomes infinitely more likely than the seller's signal to capture the true valuation θ when $\varepsilon \to 0$. This special feature of the perturbation implies that when deciding whether or not to challenge the buyer, S will privilege the information she infers from observing B's strategy over her own information. It is worth noting that the results established on the lack of existence of a truthful equilibrium in pure strategies hold for any other priors with full support. However, using the former signal structure will be useful once we move to mixed strategies. Also, for simplicity we shall keep the payments under the perturbed mechanism the same as in the above MR mechanism under common knowledge, and assume that B must participate in the mechanism. We could easily adjust the payments accordingly and assume voluntary participation.

Now, we claim that there is no perfect Bayesian (or sequential) equilibrium in pure strategies in which the buyer always reports truthfully. By way of contradiction, suppose there is such an equilibrium, and suppose that B gets signal s'_B. Then B believes that, regardless of what signal player S gets, the value of the good is greater than 10 in expectation. So B would like to announce "low" if he expects that subsequently to such an announcement, S will not challenge. Now, suppose B announces low. In a fully revealing equilibrium, S will infer that B must have seen signal s''_B if B announces low. But then, under the above signal structure, S now believes that there is a large probability that $\theta = \theta''$ and therefore S will not challenge. But then, at stage 1, anticipating that S will not challenge, B will prefer to announce "low" when he receives signal s'_B. Therefore there does not exist a truthfully revealing perfect Bayesian perfect (or sequential) equilibrium in pure strategies and consequently the above social choice function can no longer be implemented through the above MR mechanism in pure strategies.

2.3.2 Allowing for mixed strategies

The result that there are no truthful perfect Bayesian (or sequential) equilibria in pure strategies leaves open the possibility that there are mixed strategy equilibria in which the mixing probability on the truthful announcement goes to one as ε goes to zero, in the way that the pure-strategy Stackelberg equilibrium can be approximated by a mixed equilibrium of a "noisy commitment game"(van Damme and Hurkens (1997)). We show below that this is not the case.

Specifically, let σ'_B denote the probability that B announces "low" after seeing signal s'_B, and let σ''_B be the probability B announces "high" after seeing s''_B , as in the following table:

	High	Low
s'_B	$1-\sigma'_B$	σ'_B
s''_B	σ''_B	$1-\sigma''_B$

The corresponding mixing probabilities for player S are

	Challenge	Don't Challenge
s'_S	$1-\sigma'_S$	σ'_S
s''_S	σ''_S	$1-\sigma''_S$

Then for mixed strategy equilibria of the mechanism to converge to the pure information equilibrium where the buyer announces the valuation truthfully, we should have $\sigma'_B, \sigma''_B, \sigma'_S$ and σ''_S all converge to 0 as $\varepsilon \to 0$. However, this is not the case, as shown by the following:

Proposition 1. *For any fine F there is no sequence of equilibrium strategies σ_B, σ_S such that $\sigma'_B, \sigma''_B, \sigma'_S$ and σ''_S all converge to 0 as $\varepsilon \to 0$.*

Proof. We proceed by contradiction and assume that there exists a sequence of equilibrium strategies σ_B, σ_S such that $\sigma'_B, \sigma''_B, \sigma'_S$ and σ''_S all converge to 0 as $\varepsilon \to 0$**.** First, we note that under the above signal structure, the probability that $\theta = \theta''$ conditional upon (i) B receiving the low signal s''_B; (ii) B announcing the low state, i.e. playing "L"; (iii) S challenging, namely $\Pr(\theta'' \mid s''_B, L, C)$, is equal to

$$\begin{aligned}
&\Pr(\theta'' \mid s''_B, L, C) \\
&\quad = \frac{\alpha\varepsilon(1-\sigma'_S) + \alpha\left(1-\varepsilon-\varepsilon^2\right)\sigma''_S}{\left[\alpha\varepsilon + (1-\alpha)\frac{\varepsilon^2}{2}\right](1-\sigma'_S) + \left[\alpha\left(1-\varepsilon-\varepsilon^2\right) + (1-\alpha)\frac{\varepsilon^2}{2}\right]\sigma''_S} \\
&\quad = \frac{1}{1 + \frac{(1-\alpha)\frac{\varepsilon^2}{2}\left[(1-\sigma'_S)+\sigma''_S\right]}{\alpha\varepsilon(1-\sigma'_S)+\alpha(1-\varepsilon-\varepsilon^2)\sigma''_S}} = \frac{1}{1 + \frac{(1-\alpha)\frac{1}{2}\left[(1-\sigma'_S)+\sigma''_S\right]}{\frac{\alpha}{\varepsilon}(1-\sigma'_S)+\frac{\alpha}{\varepsilon^2}(1-\varepsilon-\varepsilon^2)\sigma''_S}}
\end{aligned}$$

so that $\Pr(\theta'' \mid s''_B, L, C) \longrightarrow 1$ when $\varepsilon \to 0$. Basically, this limit results follow from the fact that as $\varepsilon \to 0$, B's information prevails over S's information in the above signal structure. Similarly, the probability that $\theta = 14$ conditional upon (i) B receiving the high signal s'_B; (ii) B announcing the low state, i.e playing "L"; (iii) S challenging, namely $\Pr(\theta' \mid s'_B, L, C)$, is equal to

$$\begin{aligned}
&\Pr(\theta' \mid s'_B, L, C) \\
&\quad = \frac{(1-\alpha)\left(1-\varepsilon-\varepsilon^2\right)(1-\sigma'_S) + (1-\alpha)\varepsilon\sigma''_S}{\left[(1-\alpha)\left(1-\varepsilon-\varepsilon^2\right) + \alpha\frac{\varepsilon^2}{2}\right](1-\sigma'_S) + \left[(1-\alpha)\varepsilon + \alpha\frac{\varepsilon^2}{2}\right]\sigma''_S}
\end{aligned}$$

so that $\Pr(\theta' \mid s'_B, L, C) \to 1$. Hence, at stage 3, for ε sufficiently small, buyer B who received the high signal s'_B accepts the good at price 6 whereas buyer B who received the low signal s''_B does not accept it. In particular at stage 3, player B plays in pure strategies. Then, moving back to stage 1, the expected payoff of buyer B who received signal s''_B and plays H tends to -4 while the expected payoff from the buyer who received signal s''_B and played L tends to 0. So for ε small, there is no σ that makes player B indifferent between H and L. The same reasoning applies to buyer B who receives signal s'_B, so B plays in pure strategies in stage 1. Similar reasoning as in the pure strategy equilibrium case thus shows that σ is not an equilibrium. ■

This shows that one appealing property of the unique equilibrium in the MR mechanism under common knowledge (namely, that this is a truthful equilibrium) disappears once we introduce small common p-belief value perturbations. In the next subsection we show the non-robustness of another appealing property of the MR mechanism under common knowledge, namely that it uniquely implements any desired social choice function.

2.4 Existence of persistently bad sequential equilibria

So far we have shown that truth-telling is not a robust equilibrium outcome of the MR mechanism when allowing for common p-belief value perturbations. But in fact one can go further and exhibit arbitrarily small common p-belief value perturbations for which the above MR mechanism also has a "bad equilibrium" where the buyer reports "Low" regardless of his signal, which in turn leads to a sequential equilibrium outcome which remains bounded away from the sequential equilibrium outcome under common knowledge.

Consider the same MR mechanism as before, with the same common prior $\mu(\theta') = 1 - \alpha$ and $\mu(\theta'') = \alpha$**,** but with the following perturbation ν^ε of signals about θ:

ν^ε	s'_B, s'_S	s'_B, s''_S	s''_B, s'_S	s''_B, s''_S
θ'	$(1-\alpha)(1-\varepsilon^2)$	$(1-\alpha)\varepsilon^2/3$	$(1-\alpha)\varepsilon^2/3$	$(1-\alpha)\varepsilon^2/3$
θ''	$\alpha\varepsilon^2$	$\alpha\varepsilon/2$	$\alpha\varepsilon/2$	$\alpha(1-\varepsilon-\varepsilon^2)$

In what follows, we shall construct a sequential equilibrium of the perturbed game with prior ν^ε whose outcome differs substantially from that with complete information.

Thus, consider the following strategy profile of the game with prior ν^ε. B announces low regardless of his signal. If B has announced low, S does not challenge regardless of her signal. Off the equilibrium path, i.e. if B announced low and S subsequently challenged, then B always rejects S's offer. These are our candidate strategies for sequential equilibrium. To complete the description of the candidate sequential equilibrium, we also have to assign beliefs over states and signals for each signal of each player and for any history of play. Before playing the game but after receiving their private signals, we assume that agents' beliefs are given by ν^ε conditioned on their private signals. Similarly, if S has the opportunity to move (which in turn requires that B would have played low), we assume that her posterior beliefs are based on ν^ε together with her private signal. Finally, out of equilibrium, if B is offered the good for 6 (which requires that S will have challenged), we assume that B always believes with probability one that the state is $\theta = 10$ and that S has received signal s''_S.

So what we want to show is that for $\varepsilon > 0$ sufficiently small, the above strategy profile is *sequentially rational* given the beliefs we just described and that conversely these beliefs are *consistent* given the above strategy profile (see Kreps and Wilson (1982)).

Here we shall check sequential rationality, and then provide the basic intuition for the belief consistency part of the proof in footnote 9 below. To establish sequential rationality, we solve the game backward. At Stage 3, regardless of his signal, B believes with probability one that the state

is $\theta = 10$. Accepting S's offer at 6 generates $10 - 9 - 6 = -5$ and rejecting it generates $5 - 9 = -4$. Thus, it is optimal for B to reject the offer. Moving back to Stage 2, if S chooses "Challenge," S anticipates that with probability one her offer at 6 will be rejected by B in the next stage, thus S anticipates a payoff approximately equal to $7 - 9 = -2$ if her signal is s'_S and to $5 - 9 = -4$ if the signal is s''_S as ε becomes small. On the contrary, if S chooses "No Challenge," S guarantees a payoff of 10. Thus, regardless of her signal, it is optimal for S not to challenge. Moving back to Stage 1, B "knows" that S does not challenge regardless of her signal. Now, suppose that B receives s'_B. Then, as ε becomes small, B believes with high probability that the state is θ' so that his expected payoff approximately results in $14 - 10 = 4$. This is larger than 0, which B obtains when announcing "High." Therefore, it is optimal for B to announce "Low." Obviously, this reasoning also shows that when B has received signal s''_B, it is optimal for her to announce "Low."[9]

As we will see in the next section, the fact that the MR mechanism cannot implement even approximate truth-telling under common p-belief value perturbations is closely related to the fact that the social choice function we tried to implement is not Maskin-monotonic. But before we turn to a more general analysis of the non-robustness of subgame implementation using MR mechanisms, let us refresh the reader's memory about Maskin's monotonicity axiom and Maskin's Nash implementation theorem, and also explain why the social choice function we try to implement in this Hart-Moore example is not Maskin-monotonic.

2.5 This example does not satisfy Maskin-monotonicity

2.5.1 Maskin's Nash implementation theorem

Recall that a social choice function f on a payoff relevant state space Θ is Maskin-monotonic if for all pair of states of nature (preference profiles) θ and θ', if $x = f(\theta)$, and no individual ranks x lower when moving from θ to θ', then $x = f(\theta')$. Maskin's Nash implementation theorem says that if f is Nash-implementable (that is, there exists a mechanism $\Gamma = (M, g)$ where $m = (m_1, \ldots, m_n) \in M = M_1 \times \cdots \times M_n$ denotes a strategy profile and $g : M \to A$ is the outcome function (which maps strategies into outcomes), *and if* for all θ the Nash Equilibrium outcome of that mechanism in state θ is precisely $f(\theta)$, then f must be Maskin monotonic.

Let us summarize the proof, which we shall refer to again below. By way of contradiction, if f were not monotonic, then (if u_i denote player i's utility function) there would exist θ and θ' such

[9] To establish belief consistency, we need to find a sequence of totally mixed strategies that converges toward the pure strategies described above and so that beliefs obtained by Bayes rule along this sequence also converge toward the beliefs describe above. It is easy to see that under any sequence of totally mixed strategies converging toward the pure strategies described above, the induced sequence of beliefs about θ will converge toward ν^ε conditioned on private signals along the equilibrium path of the pure-strategy equilibrium. When B is offered the good at 6, S has deviated from the equilibrium path due to the "trembles." Beliefs about θ are then determined by the relative probability that S has trembles after the different signals. For instance, if one chooses a sequence of totally mixed strategies under which it becomes infinitely more likely that S has trembled after receiving s''_S rather than when receiving s'_S, then B will assign probability one to S receiving signal s''_S.

that for all players i and for all alternative b

$$u_i(f(\theta);\theta) \geq u_i(b;\theta) \Longrightarrow u_i(f(\theta);\theta') \geq u_i(b;\theta') \tag{I}$$

and nevertheless $f(\theta) \neq f(\theta')$. But at the same time if f is Nash-implementable there exists a mechanism $\Gamma = (M, g)$ such that $f(\theta) = g(m_\theta^*)$ for some Nash equilibrium m_θ^* of the game $\Gamma(\theta)$. By definition of Nash equilibrium, we must have

$$u_i(f(\theta);\theta) = u_i(g(m_\theta^*);\theta) \geq u_i(g(m_i, m_{-i}^*);\theta), \forall m_i.$$

But then, from (I) we must also have

$$u_i(f(\theta),\theta') = u_i(g(m_\theta^*);\theta') \geq u_i(g(m_i, m_{-i}^*);\theta'), \forall m_i,$$

so that $f(\theta)$ is also a Nash equilibrium outcome in state θ'. But then if the mechanism implements f, we must have $f(\theta) = f(\theta')$, a contradiction.

2.5.2 The social choice function in Hart-Moore is non monotonic

It is easy to show that the social choice function in this Hart-Moore example is not Maskin monotonic. The set of social outcomes (or alternatives) A is defined as

$$A = \left\{(q, y_B, y_S) \in [0,1] \times \mathbb{R}^2 \text{ such that } y_B + y_S \leq 0\right\},$$

where q is the probability that the good is traded from S to B, and y_B, y_S are the transfers of B and S respectively. Their sum must be non-positive, i.e. we allow for penalties paid to a third party.

We have two states of the world θ' and θ'', which correspond respectively to the good being of high and of low quality, and we have just seen that a SCF $f : \Theta \to A$ which is Maskin-monotonic, must satisfy:

$$f\left(\theta''\right) = f\left(\theta'\right).$$

At the same time, the social choice function we seek to implement in this Hart-Moore example requires that

$$f\left(\theta''\right) = (1, -10, 10),$$
$$f\left(\theta'\right) = (1, -14, 14).$$

Clearly $f\left(\theta''\right) \neq f\left(\theta'\right)$, but the buyer ranks outcome (1,-10,10) at least as high under θ' as under θ'', while the seller has the same preferences in the two states. Thus, f is not Maskin-monotonic, so Maskin's theorem implies that this f is not Nash-implementable. It is Moore-Repullo (MR) implementable under common knowledge, but it is not MR implementable under common p-belief

value perturbations.

Our analysis in the next two sections is motivated by the following two questions: (1) Is the nonexistence of truth-telling equilibria in arbitrarily small common p-belief value perturbations of the above MR mechanism linked to the SCF f being non Maskin-monotonic? (2) Is the existence of a sequence of bad sequential equilibrium in arbitrarily small common p-belief value perturbations of the above MR mechanism, directly linked to f being non Maskin-monotonic?

In the next section we shall consider a more general version of the MR mechanism and then link the failure of MR mechanisms to implement truth-telling in equilibrium under common p-belief value perturbations to the non-monotonicity of the corresponding SCF. Then in Section 4 we will consider any sequential mechanism that implements a non-monotonic SCF under common knowledge, and show that for an arbitrarily small common p-belief value perturbation of the game defined by the mechanism under common knowledge there exists a bad sequential equilibrium whose outcome remains bounded away from the good equilibrium outcome under common knowledge**,** even when the size of the perturbation tends to zero.

3 More general Moore-Repullo mechanisms

Moore and Repullo (1988) consider a more general class of extensive form mechanisms, which we shall refer to as "MR mechanisms". Under complete information, these mechanisms work well in fairly general environments. They also outline a substantially simpler mechanism which yields truth telling in environments where there is transferable utility. Since this is the most hospitable environment for subgame perfect implementation, and because most contracting settings are in economies with money, we shall focus on it.

3.1 Setup

Let there be two players 1 and 2, whose preferences over a social decision $d \in D$ are given by $(\theta_1, \theta_2) \in \Theta_1 \times \Theta_2 = \Theta$ where $\Theta_i = \left\{\theta_i^1, ..., \theta_i^n\right\}$ for each $i = 1, 2$.[10]The players have utility functions

$$u_1((d, t_1, t_2), \theta_1) = U_1(d; \theta_1) - t_1$$

and

$$u_2((d, t_1, t_2), \theta_2) = U_2(d; \theta_2) + t_2,$$

where d is a collective decision, t_1 and t_2 are monetary transfers. Preference characteristics (θ_1, θ_2) are common knowledge among the two parties, but not verifiable by a third party.

Let $f = (D, T_1, T_2)$ be a social choice function where for each $(\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ the social decision is $d = D(\theta_1, \theta_2)$ and the transfers are $(t_1, t_2) = (T_1(\theta_1, \theta_2), T_2(\theta_1, \theta_2))$.

[10]Moore and Repullo (1988) allow for an infinite state space but impose bounds on the utility functions which are automatically satisfied in the finite case.

Moore and Repullo (1988) propose the following class of mechanisms. The mechanisms involve two phases, where phase i is designed so as to elicit truthful revelation of θ_i. Each phase in turn consists of three stages. The game begins with phase 1, in which player 1 announces a value θ_1 which we now outline, and then carries on with phase 2 in which player 2 announces θ_2.

1. Player 1 announces a preference θ_1, and we proceed to stage 2.

2. If player 2 agrees with player 1's announcement, then phase 1 ends and we proceed to phase 2. If player 2 does not agree and "challenges" player 1 by announcing some $\phi_1 \neq \theta_1$, then we proceed to stage 3.

3. Player 1 chooses between
$$\{x; t_x + \Delta\}$$
and
$$\{y; t_y + \Delta\},$$
where these functions are specified by the mechanism such that
$$U_1(x; \theta_1) - t_x > U_1(y; \theta_1) - t_y$$
and
$$U_1(x; \phi_1) - t_x < U_1(y; \phi_1) - t_y.$$
If player 1 chooses $\{x; t_x + \Delta\}$, which proves player 2 wrong in his challenge (in the Hart-Moore example above, this corresponds to the buyer refusing the offer at price 6), then player 2 receives $t_2 = t_x - \Delta$ and a third party receives 2Δ. However, if player 1 chooses $\{y; t_y + \Delta\}$, which confirms player 2's challenge (in the above Hart-Moore example, this corresponds to the buyer taking up the offer at price 6), then player 2 receives $t_2 = t_y + \Delta$.

Phase 2 is the same as phase 1 with the roles of players 1 and 2 reversed, i.e. with player 2 announcing θ_2 in the first stage of that second phase. We use the notation stage 1.2, for example, to refer to phase 1, stage 2.

The Moore-Repullo logic then works as follows when the state of nature θ is common knowledge. If player 1 lies at stage 1.1 then player 2 will challenge her and then at stage 1.3 player 1 will find it optimal to choose $\{y; t_y + \Delta\}$. If Δ is sufficiently large then at stage 1, anticipating player 2's subsequent challenge, player player 1 will find it optimal to announce the truth, and thereby implement the choice function f. Moreover, player 2 will be happy with receiving $t_y + \Delta$. If player 1 tells the truth at stage 1.1 then player 2 will not challenge because she knows that player 1 will choose $\{x; t_x + \Delta\}$ at stage 1.3 which will cause player 2 to pay the fine of Δ.

3.2 Perturbing the information structure

We now show that this result does not hold for small common p-belief value perturbations of the information structure. We consider the following type of information structure. Each agent $i = 1, 2$ receives a signal $s_i^{k,l}$ where k and l are both integers in $\{1, ..., n\}$; the set of signals of player i is denoted S_i. We assume that the prior joint probability distribution ν^ε over the product of signal pairs and state of nature is such that, for each (k, l) :

$$\nu^\varepsilon(s_1^{k,l}, s_2^{k,l}, \theta_1^k, \theta_2^l) = \mu(\theta_1^k, \theta_2^l)[1 - \varepsilon - \varepsilon^2]$$

$$\nu^\varepsilon(s_1^{k,l_1}, s_2^{k_2,l}, \theta_1^k, \theta_2^l) = \mu(\theta_1^k, \theta_2^l)\frac{\varepsilon}{n^2 - 1} \text{ for } (k_2, l_1) \neq (k, l)$$

$$\nu^\varepsilon(s_1^{k_1,l_1}, s_2^{k_2,l_2}, \theta_1^k, \theta_2^l) = \mu(\theta_1^k, \theta_2^l)\frac{\varepsilon^2}{n^4 - n^2} \text{ for } k_1 \neq k \text{ or } l_2 \neq l$$

where μ is a complete information prior over states of nature, i.e. a prior satisfying $\mu(s_1^{k_1,l_1}, s_2^{k_2,l_2}, \theta_1^k, \theta_2^l) = 0$ whenever $(k_i, l_i) \neq (k, l)$ for some player i. This corresponds to a common p-belief perturbation such that each player $i's$ signal is more informative about his own preferences than those of the other player.

Denote the probability that player 1 announces θ_1^j conditional on seeing signal $s_1^{k,l}$ as $\sigma_{k,l}^j$. Similarly let the probability that player 2 announces θ_1^j (at stage 2) conditional on observing signal $s_2^{k,l}$ be $\mu_{k,l}^j$ (w.l.o.g. this is not conditioned upon player 1's move). In the second phase of the mechanism (designed to elicit player 2's preferences) the corresponding mixing probabilities are as follows. The probability that player 2 announces θ_2^j conditional on seeing signal $s_2^{k,l}$ is $\rho_{k,l}^j$ and that probability the player 1 announces θ_2^j (at stage 2) conditional on observing signal $s_1^{k,l}$ is $\tau_{k,l}^j$. Here again, we can ignore dependence with respect to past histories of moves.

First, when restricting attention to pure strategy equilibria, a fundamental result relates the non-robustness of truth-telling in the Moore-Repullo mechanisms to the failure of Maskin monotonicity of the social choice function f. More specifically, we introduce the following:

Definition 1. *A SCF f is **strategy-proof** if for each player i and each θ_i,*

$$u_i(f(\theta_i, \theta_{-i}), \theta_i) \geq u_i(f(\theta_i', \theta_{-i}), \theta_i) \text{ for all } \theta_i' \text{ and } \theta_{-i}.$$

In other words, a social choice function is strategy-proof if it is implementable in dominant strategies through a direct mechanism whereby the players are asked to announce their preference parameter. Below we shall argue that strategy proofness implies Maskin monotonicity under very weak assumptions on players' preferences. Under such assumptions, the non-robustness of the MR mechanism to common p-belief value perturbations, follows immediately from the following.

Theorem 1. *Suppose that a non-strategy proof SCF f is MR implementable under common knowledge. Fix any complete information prior μ. There exists a sequence of priors $\{\nu^\varepsilon\}_{\varepsilon>0}$ that converges to the complete information prior μ such that there is no pure equilibrium strategies under*

which player 1 tells the truth in phase 1 and player 2 tells the truth in phase 2 i.e. so that $\sigma_{k,l}^k = 1$ *for all* k *and* l *and* $\rho_{k,l}^l = 1$ *for all* k *and* l*.*

Thus, if f is a non-strategy proof SCF which is implemented by a MR mechanism under common knowledge, truth-telling by the first player in each phase cannot be sequentially rational for beliefs consistent with Bayes rule. This implies that there is no perfect Bayesian equilibrium of this mechanism in which player i tells the truth in phase i and is robust to small p-belief value perturbations. This reasoning also implies the non-existence of a good pure equilibrium. The following proof is quite intuitive:

Proof. First, consider the same common p-belief perturbation ν^ε as specified before, whereby if player 2 sees that player 1's announcement about θ_1 is different from her signal, she tends to disregard her own information on Θ_1 and follow player 1's announcement (and symmetrically for player 1 vis a vis player 2 regarding signals over Θ_2).

Now, suppose that f is not strategy-proof. Then there exist some player, say player 1, and states $\theta_1^k, \theta_1^{k'}, \theta_2^l$ such that

$$u_1(f(\theta_1^k, \theta_2^l), \theta_1^k) < u_1(f(\theta_1^{k'}, \theta_2^l), \theta_1^k).$$

Then, we claim that there is no pure strategy equilibrium in which player 1 reports truthfully in phase 1 and player 2 reports truthfully in phase 2. By way of contradiction, suppose there is such an equilibrium, and suppose that player 1 gets signal $s_1^{k,l}$ whereas player 2 gets signal $s_2^{k,l}$. Yet, player 1 would like to announce "$\theta_1^{k'}$" if she expects that subsequently to such an announcement, player 2 announces "$\theta_1^{k'}$" as well and then tells the truth in phase 2 so that the outcome is $f(\theta_1^{k'}, \theta_2^l)$. But this is precisely what will happen: In a fully revealing equilibrium, player 2 will infer that player 1 must have seen a $s_1^{k',\tilde{l}}$-type signal, therefore player 2 will believe with high probability that the state must be $(\theta_1^{k'}, \theta_2^l)$. Consequently, player 2 will not challenge player 1's announcement. But then, anticipating this, player 1 will announce "$\theta_1^{k'}$" and thereby receive $f(\theta_1^{k'}, \theta_2^l)$ instead of $f(\theta_1^k, \theta_2^l)$. This in turn shows that there does not exist a truthfully revealing equilibrium in pure strategies. ■

Theorem 1 links the non-robustness of the MR mechanism to the failure of Maskin monotonicity of the social choice function to be implemented by that mechanism. This follows from the above result and from the fact that strategy-proofness implies a weak version of Maskin monotonicity, namely, that for any θ, θ' such that

$$\forall i \in N \text{ and } \forall b \in A : u_i(f(\theta); \theta_i) \geq u_i(b; \theta_i) \Rightarrow u_i(f(\theta); \theta_i') > u_i(b; \theta_i')$$

we have $f(\theta) = f(\theta')$. Strategy-proofness also implies the usual Maskin monotonicity condition when preferences over outcomes in $f(\Theta)$ are strict.[11] For instance, in the Hart-Moore example in

[11] To see that strategy proofness implies the weak monotonicity condition, note that if $f(\theta) \neq f(\theta')$, it must be that there is some player i and some $\hat{\theta}_{-i}$ such that $f(\theta_i, \theta_{-i}) = f(\theta_i, \hat{\theta}_{-i}) \neq f(\theta_i', \hat{\theta}_{-i})$, and so in particular $\theta_i \neq \theta_i'$. Hence, strategy-proofness of f implies that for this player i, $u_i(f(\theta_i, \theta_{-i}); \theta_i) = u_i(f(\theta_i, \hat{\theta}_{-i}); \theta_i) \geq u_i(f(\theta_i', \hat{\theta}_{-i}); \theta_i)$ and $u_i(f(\theta_i, \theta_{-i}); \theta_i) = u_i(f(\theta_i, \hat{\theta}_{-i}), \theta_i') \leq u_i(f(\theta_i', \hat{\theta}_{-i}); \theta_i')$, and setting $b = f(\theta_i', \hat{\theta}_{-i})$ yields the weak monotonicity con-

Section 2, the social choice function is not monotonic and preferences over $f(\Theta)$ are strict, so the social choice function in that example is not strategy-proof.

Note that the above result does not preclude the existence of mixed strategy equilibria where truth-telling by one or two players in each phase is robust to small p-belief value perturbations. Moreover, the above result provides a necessary condition for the robustness of truth-telling by player i in phase i, without requiring truth-telling by player j as well. If we allow for mixed strategies and require that both players tell (at least, approximately) the truth in each of the two phases, then *no* social choice function $f = (D, T_1, T_2)$ can be implemented by the general MR mechanism in such a way that truth-telling by both players in each phase, is a sequential equilibrium outcome which is robust to common p-belief value perturbations. More formally, in the Appendix we prove the following:[12]

Theorem 2. *Suppose that a SCF f is MR implementable under common knowledge. Fix any complete information prior μ. There exists a sequence of priors $\{\nu^\varepsilon\}_{\varepsilon>0}$ that converges to the complete information prior μ such that there is no sequence of sequential equilibrium strategy profiles that converges to truth-telling, i.e. so that $\sigma^j_{k,l}, \mu^j_{k,l}$ converge to 0 as $\varepsilon \to 0$ for all $k \neq j$ and all l, and $\rho^j_{k,l}$,$\tau^j_{k,l}$ converge to 0 as $\varepsilon \to 0$ for all $l \neq j$ and all k.*

Let us make two remarks at this stage. First, the non-robustness of truth-telling as a sequential equilibrium outcome of the MR mechanism, is of interest both because truth-telling is cognitively simple and because the non-existence of a truthful sequential equilibrium here implies the non-existence of a desirable pure equilibrium, and implementation theory has focused on pure-strategy equilibrium. Second, neither of the non-robustness results of this section rule out the possibility that some SCF f be implemented as the limit of some mixed (non-truthful) sequential equilibrium outcomes. However, in the next section we will show that if f is not Maskin-monotonic but yet can be implemented by the MR or by any other extensive form mechanism under common knowledge, then there always exist arbitrarily small common p-belief value perturbations under which there also exist sequential equilibria with undesirable outcomes**.**

4 Any mechanism

In this section we consider the set of all extensive form mechanisms, and show that whenever the social choice function to be implemented is not Maskin-monotonic, then whenever there exists a

dition. Finally, note that if preferences over outcomes in $f(\Theta)$ are strict then $u_i(f(\theta_i, \theta_{-i}), \theta'_i) = u_i(f(\theta_i, \hat{\theta}_{-i}), \theta'_i) < u_i(f(\theta'_i, \hat{\theta}_{-i}); \theta'_i)$ and so the above argument yields the usual Maskin-monotonicity condition.

[12] To gain intuition for why requiring approximate truth-telling by *both* players in each phase precludes robust implementation of *any* SCF by the MR mechanism, suppose that both players receive a signal which is highly correlated with the true state. Player 1 plays first in phase 1, so if player 1 announces a signal that is highly correlated with some state $\widehat{\theta}$, then player 2 (playing in second in phase 1) will believe that player 1 has told the truth (because by assumption player 1's announcement is close to truthful). But the mechanism is built in such a way that player 2 never wants to challenge player 1 if she thinks that player 1 is telling the truth (otherwise at stage 3 player 2 will be punished), so player 2, if she is not challenging, will also announce $\widehat{\theta}$ and so will not follow her private signal and thus she is not reporting truthfully.

"good" sequential equilibrium for a mechanism that implements it under common knowledge, there always also exist a "bad" sequential equilibrium in arbitrarily small p-belief value perturbations of that mechanism. We begin by presenting the argument in a nutshell, using the Hart-Moore example to illustrate our point. We then introduce a non-MR mechanism in the context of the Hart-Moore example, which implements truth-telling even under a particular common p-belief perturbation, and yet we show that again in that case a bad equilibrium can be constructed whenever $\varepsilon > 0$. Finally, we proceed to state and establish the more general result.

4.1 Overview of the main result

In this subsection we state the main result and provide the reader with an intuition for the proof. The main idea is that by introducing just a small amount of incomplete information, one can rapidly increase the sets of (sequential) beliefs that are consistent with Bayesian rationality. As a result, one can turn a bad (non-sequential) Nash equilibrium of an extensive-form mechanism that implements a non-monotonic SCF f under common knowledge into a sequential equilibrium of the perturbed game.

More specifically, suppose there are n players where each player i has a utility function $u_i(a, \theta)$, over outcomes (or alternatives) $a \in A$. In the perturbations we consider, players do not observe the state of nature θ directly, but are informed about it through private signals. An extensive form mechanism Γ together with a state $\theta \in \Theta$ defines an extensive form game $\Gamma(\theta)$, and let $SPE(\Gamma(\theta))$ denote the set of subgame perfect equilibria of the game $\Gamma(\theta)$. Here is an informal statement of the main result:

Result: Assume finite state space and finite strategy spaces.[13] And suppose that a mechanism Γ subgame perfect implements a non-Maskin monotonic SCF f under common knowledge. Then there exists a sequence of common p-belief value perturbations parametrized by some ε and a corresponding sequence of sequential equilibria of the games induced by Γ under this sequence of perturbations, whose outcomes under $\Gamma(\theta)$ do not converge to $f(\theta)$ in some state θ as $\varepsilon \to 0$.

In particular, under the usual additional conditions under which Maskin-monotonicity is sufficient for Nash-implementation, this result implies that whenever a SCF cannot be implemented using static mechanisms (with Nash equilibrium as the solution concept), there is no hope of implementing it using sequential mechanisms if we want such mechanisms to be robust to common p-belief value perturbations**.**

Intuition for the proof: Suppose that the SCF f is not Maskin-monotonic. Then there exist θ' and θ'' such that for all players $i \in N$ and for all alternative $b \in A$

$$u_i(f(\theta'); \theta') \geq u_i(b; \theta') \Longrightarrow u_i(f(\theta'); \theta'') \geq u_i(b; \theta'') \tag{I}$$

and nevertheless $f(\theta') \neq f(\theta'')$. At the same time, since the extensive-form mechanism Γ implements f, there exists a subgame perfect equilibrium (SPE) $m_{\theta'} \in SPE(\Gamma(\theta'))$ such that $g(m_{\theta'}) = f(\theta')$.

[13] In the Appendix we extend the result to the case of countable strategy sets.

But then using the same argument as in the proof of Maskin's theorem summarized in Section 2 above, $m_{\theta'}$ is also a Nash equilibrium in state θ'', and necessarily a "bad" Nash equilibrium since $f(\theta') \neq f(\theta'')$.

The remaining part of the proof follows from the fact that one can use common p-belief value perturbations to "rationalize" this bad Nash equilibrium and turn it into a sequential equilibrium of the perturbed games, in the same way as the construction in Section 2 above that showed the non-robustness of the particular MR mechanism considered in that section.

As a concrete example, consider again the MR mechanism studied in Section 2. Under common knowledge, having B always announce θ'' at stage 1 and then having S never challenge at stage 2, is a bad Nash equilibrium which is not a sequential equilibrium under common knowledge. In particular, if stage 3 were to be reached under common knowledge, then B would just infer that S deviated from the equilibrium, but never update his beliefs about the true valuation θ or about S's perception of θ.

However, perturbing the signals about θ changes the picture radically. Now, if stage 3 is reached, then B updates his beliefs about which signal S might have seen. In particular, if B's updating puts enough weight on S having the low signal, then B will not take the offer at price 6; then, anticipating this at stage 2, S will indeed not challenge in equilibrium. Note that by perturbing the signal structure we have enlarged the set of consistent beliefs: under common knowledge it could not be a consistent belief that S saw θ'' if B "knew" that the state was θ', but this can be consistent under the perturbation. This is the key to how the perturbation turns a bad (non-sequential) Nash equilibrium of the game with common knowledge into a sequential equilibrium.

4.2 A more formal statement of the main result

Now, we move from intuition and examples to the formal statement of the result, and refer the reader to the Appendix for the formal proof.

4.2.1 The environment

In what follows, we consider a more general environment, with is a finite set $N = \{1, ..., n\}$ of players, with $n \geq 2$, and a set A of social alternatives, or outcomes. From now on, we no longer assume that agents have quasilinear preferences with transferable money, as was needed for MR mechanisms. Each player i has a utility function $u_i : A \times \Theta \to \mathbb{R}$, where Θ is a finite set of states of nature.[14] Players do not observe the state directly, but are informed of the state via signals. Player i's signal set is S_i which, for simplicity, we identify with Θ. A signal profile is an element $s = (s_1, ..., s_n) \in S \equiv \times_{i \in N} S_i$. When the realized signal profile is s, each player i observes only his own signal s_i. We let μ denote the prior probability over $\Theta \times S$. We note $\mu(\cdot|s_i)$ for the probability measure over $\Theta \times S$ conditional on s_i. Let s^θ be the signal profile in which each player's signal is θ. *Complete information* refers to the environments in which $\mu(\theta, s) = 0$ whenever $s \neq s^\theta$ (μ will

[14] We can always interpret a partition over Θ as corresponding to a particular player i's set of types Θ_i. Thus indeed the set up considered in the previous sections is a special case of that analyzed in this section.

be then referred to as a complete information prior). Under complete information, the state, and hence the full profile of preferences, is always common knowledge among players.

We will assume for each i and θ, the marginal distribution on i's signals places strictly positive weight on every signal in every state, i.e. $\mu(s_i^\theta) \equiv [\text{marg}_{S_i}\mu](s_i^\theta) > 0$, so that Bayes rule is well-defined. Note that in case μ is a complete information prior, this implies in particular that for each $(\theta, s^\theta) \in \Theta \times S : \mu(\theta, s^\theta) > 0$.

A *social choice function* (SCF) is a mapping $f : \Theta \to A$. A *mechanism* is an extensive game form $\Gamma = (\mathcal{H}, M, \mathcal{Z}, g)$ where (1) $\mathcal{H}$ is the set of all histories; (2) $M = M_1 \times \cdots \times M_n$, $M_i = \times_{h \in \mathcal{H}} M_i(h)$ for all i where $M_i(h)$ denotes the set of available messages for i at history h; (3) $\mathcal{Z}$ describes the history that immediately follows history h given that the strategy profile m has been played; and (4) g is the outcome function that maps the set of terminal histories H_T into A. From the outset, we clarify the class of mechanisms to be considered. First, we restrict our multi-stage games with observed actions by assuming that at each history h, all players know the entire history of the play, and if more than one player moves at h, they do so simultaneously.[15] Second, we assume that the mechanism has a finite number of stages.

The following notation will be useful: An element of $M(h) = M_1(h) \times \cdots \times M_n(h)$, say $m(h) = (m_1(h), ..., m_n(h))$ is a message profile at h while $m_i(h)$ is i's message at h. If $\#M_i(h) > 1$ and $\#M_j(h) > 1$ then players i and j move simultaneously after history h, whereas if $\#M_i(h) > 1$ and $\#M_j(h) = 1$ for all $j \neq i$ then player i is the only one to move. Histories and messages are tied together by the property that $M(h) = \{m : (h, m) \in \mathcal{H}\}$. An element of M_i is a pure strategy; and an element of M is a pure strategy profile. We sometimes write $m \mid_h = (m_1 \mid_h, ..., m_n \mid_h)$ for the profile of pure strategies starting from history h.

There is an initial history $\emptyset \in \mathcal{H}$, and $h_t =: (\emptyset, m^1, m^2, ..., m^{t-1})$ is the history at the end of period t, where for each k, $m^k \in M(h_k)$. If for $t' \geq t+1, h_{t'} = (h_t, m^t, ..., m^{t'-1})$, then $h_{t'}$ follows history h_t. As Γ contains finitely many stages, there is a set of terminal histories[16] $H_T \subset \mathcal{H}$ such that $H_T = \{h \in \mathcal{H} : \text{there is no } h' \text{ following } h\}$. Given any strategy profile m and any history h, there is a unique terminal history denoted by $h_T[m, h]$. Formally, let $\mathcal{Z} : M \times \mathcal{H} \to \mathcal{H}$ be the mapping where

$$\mathcal{Z}[m, h] = \begin{cases} (h, m(h)) \text{ if } h \notin H_T \\ h \text{ otherwise} \end{cases}$$

is the history that immediately follows h whenever possible given that strategy profile m has been played; and so $h_T[m, h] = \lim_{k \to \infty} \mathcal{Z}^k[m, h]$ where $\mathcal{Z}^k[m, h] = \mathcal{Z}[m, \mathcal{Z}^{k-1}[m, h]]$. Finally, the outcome function $g : H_T \to A$ specifies an outcome for each terminal history. We will also denote $g(m; h)$ the outcome that obtains when players use strategy profile m starting from history h i.e. $g(m; h) = g(h_T[m, h])$. In what follows, we only consider finite mechanisms:

Assumption A1. *$M_i(h)$ is finite for each i and h.*

[15] This includes games of perfect information (sequential and observed moves) as a special case.

[16] Note that $M(h) = \{m : (h, m) \in \mathcal{H}\} = \emptyset$ for any $h \in H_T$.

Remark 1. *This assumption is useful when using sequential equilibrium and avoids technical complications due to the use of countably infinite (or uncountable) spaces. In the appendix, we provide additional assumptions on the class of mechanisms so that our result can be extended to countable message spaces. We believe that this extension is important because the literature often uses integer games (i.e., game where one dimension of the message is space is the set of positive integers) as part of implementing mechanisms. Moreover, we do not believe that our results critically depend on the countability assumption. We refer the reader to Duggan (1997) for the treatment of general (uncountable) message spaces. It is also worthy to note that our results would hold for arbitrary mechanisms if we used perfect Bayesian equilibrium instead of sequential equilibrium.*

A mechanism Γ together with a state $\theta \in \Theta$ defines an extensive game $\Gamma(\theta)$. A (pure strategy) Nash equilibrium for the complete information game $\Gamma(\theta)$ is an element $m^* \in M$ such that, for each player i, $u_i(g(m^*;\emptyset));\theta) \geq u_i(g((m_i, m^*_{-i});\emptyset);\theta)$ for all $m_i \in M_i$. A (pure strategy) subgame perfect equilibrium for the game $\Gamma(\theta)$ is an element $m^* \in M$ such that, for each player i, $u_i(g(m^*;h);\theta) \geq u_i(g((m_i, m^*_{-i});h);\theta)$ for all $m_i \in M_i$ and all $h \in \mathcal{H}\backslash H_T$. Let $SPE(\Gamma(\theta))$ denote the set of subgame perfect equilibria of the game $\Gamma(\theta)$. Let also $NE(\Gamma(\theta))$ denote the set of Nash equilibria of the game $\Gamma(\theta)$. We say that a mechanism implements an SCC $\mathcal{F}$ in subgame perfect equilibrium, or simply SPE-implements $\mathcal{F}$, if for each $(\theta, s^\theta) \in \Theta \times S$, we have $g(SPE(\Gamma(\theta));\emptyset) = \mathcal{F}(\theta)$.

Given a prior μ, the mechanism determines a Bayesian game $\Gamma(\mu)$ in which each player's type is his signal, and after observing his signal, player i selects a (pure) strategy from the set M_i. A strategy profile $\sigma = (\sigma_1, ..., \sigma_n)$ lists a strategy for each player i where $\sigma_i : S_i \to M_i$ and $\sigma_i(h_t, s_i)$ is the message in $M_i(h_t)$ given history h_t and signal s_i. Alternatively, we will sometimes let σ_i be a (mixed) behavior strategy i.e. a function that maps the set of possible histories and signals into the set of probability distributions over messages: $\sigma_i(\cdot \mid h_t, s_i) \in \Delta(M_i(h_t))$ is the probability distribution over $M_i(h_t)$ given history h_t and signal s_i.[17]

4.2.2 The existence of a bad sequential equilibrium with almost-perfect information

Henceforth, we assume that A is an arbitrary topological space, and we restrict our attention to SCF; our results are easily extended to the case of correspondences. We are now in a position to provide a more formal statement of our main theorem.

Theorem 3. *Assume A1. Suppose that a mechanism Γ SPE-implements a non-monotonic SCF f. Fix any complete information prior μ. There exists a sequence of priors $\{\nu^\varepsilon\}_{\varepsilon>0}$ that converges to a complete information prior μ and a corresponding sequence of sequential equilibria $\{(\phi^\varepsilon, \sigma^\varepsilon)\}_{\varepsilon>0}$ such that as ε tends to 0, $g(\sigma^\varepsilon(s^\theta);\emptyset) \nrightarrow f(\theta)$ for some $\theta \in \Theta$.*

Proof. See Appendix. ■

[17] In fact, we can show that all the results can be extended to much more general representations for preferences under uncertainty. Some additional assumptions we need for the result are very similar to the ones Chung and Ely (2003) use. The interested reader is referred to Kunimoto and Tercieux (2009) for details.

Remark 2. *The above theorem shows that, under the usual conditions ensuring that Maskin-monotonicity is sufficient for Nash-implementation, whenever a SCF cannot be implemented using static mechanisms, this SCF cannot be implemented using a mechanism that is robust to the introduction of a small amount of incomplete information.*

Remark 3. *While non-monotonic SCFs cannot be robustly implemented, things are quite different for Maskin-monotonic SCFs. For instance, using Maskin's (1999) canonical mechanism, Kunimoto (2010) shows that under suitable conditions*[18] *"any" Maskin monotonic SCF is Nash implementable by a static mechanism in a robust way: namely, for any common p-belief value perturbation, there exist at least one desirable Nash equilibrium and, moreover, no undesirable Nash equilibrium appears.*

5 Conclusion

We conclude by making a few additional remarks. First, the bad sequential in the previous section equilibria survive a standard equilibrium selection criterion. Cho (1987) defines *forward induction equilibrium,* which is an extension of the Cho-Kreps intuitive criterion to general games. The key restriction in this equilibrium concept is that the belief system assigns probability 0 to nodes in some information set h if this node can be achieved only by "bad" deviations, provided that other nodes in h can be reached by non-bad deviations. Here, "bad deviations" are deviations with the following property: Suppose that at any information set which the deviating player can reach by deviating, players are playing best-responses against some arbitrary belief that is consistent with that information set being reached. Then the deviation makes the deviating player strictly worse off compared to his equilibrium payoff. In the HM example developed in Section 2, we can show that challenging is never a bad deviation for the seller. To see this, note that when deviating to "Challenge," the seller may think that an information set under which B believes that the state θ' may occur with positive probability. Thus we can always pick an appropriate belief (for instance, one that would assign probability 1 to θ') under which it is a best reply for B to accept S's offer if S challenges. But we know that in such a case challenging for the seller makes her strictly better off compared to the equilibrium, proving that challenging cannot be a bad deviation.

A second remark concerns the implications of our analysis for the hold-up problem. For example, if we go back to the Hart-Moore example developed in Section 2 and introduce a stage zero, where S chooses investment i at cost $c(i)$, knowing that $\Pr(\theta') = \alpha i$.[19] The first-best benchmark involves maximizing total surplus from this investment, namely:

$$\max_i \{\alpha i 14 + (1 - \alpha i)10 - c(i)\}.$$

[18] More precisely, when: (1) there are at least three agents; (2) the set of outcomes is a separable space; (3) the SCF satisfies no-veto-power.

[19] We assume that $c(i)$ satisfies the following properties: $c : [0, 1/\alpha] \to \mathbb{R}_+$; $c(0) = 0$; $c(1/\alpha) = \infty$; $c' > 0$; $c'' > 0$; $c'(0) = 0$; and $c'(1/\alpha) = \infty$.

The first-order condition is

$$4\alpha = c'(i).$$

MR mechanisms under common knowledge (in particular the mechanism considered in Section 2), would implement this first-best investment by ensuring that S gets paid 14 when $\theta = \theta' = 14$and not more than 10 when $\theta = \theta'' = 10$. However, our analysis in the previous section implies that the SCF $f(\theta) = (1, -\theta, \theta)$ is not Maskin-monotonic, there is a sequence of perturbations ν_ε and a corresponding sequence of sequential equilibria for the MR mechanisms under these perturbations, where the buyer announces a low state after having observed a high signal. This in turn implies that S's payoffs do not depend on the true state and hence, S has no incentive to invest at all. In other words, subgame perfect implementation does not achieve the first best in a way which is robust to small p-belief perturbations and hence does not entirely solve the hold-up problem.

Our third remark is that the non-robustness of subgame perfect implementation does not mean that implementation is hopeless but rather suggests that we should further explore the implications of Nash implementation. We saw that the social choice function f we sought to implement in this Hart-Moore example was not Maskin-monotonic since $f\left(\theta''\right) = (1, -10, 10) \neq f\left(\theta'\right) = (1, -14, 14)$, and therefore not Nash-implementable. However, the ε-approximation of that SCF defined by

$$f^\varepsilon\left(\theta''\right) = (1-\varepsilon, -10, 10) \neq f\left(\theta'\right) = (1-\varepsilon, -14, 14),$$

is Maskin-monotonic since for example B strictly prefers $(1-\varepsilon, -10, 10)$ to $(1, -10-11\varepsilon, 10)$ when $\theta = 10$ but the reverse is true when $\theta = 14$. Therefore f^ε is Nash-implementable. For example, Abreu and Sen (1991) and Matsushima (1988) show that if agents are expected utility maximizers, then for any $\varepsilon > 0$ and any SCF f, there exists a SCF f^ε which is Maskin monotonic and is ε-close to f.[20] Hence, if f is not Maskin-monotonic and therefore not Nash implementable, we can find an ε-close SCF that is Maskin-monotonic and therefore Nash-implementable for instance in Moore and Repullo's setting[21].

A first avenue for future work would be to use our framework and the notion of robustness to common p-belief value perturbations to revisit the role of asset ownership and vertical integration in mitigating ex-ante investment and ex-post trade inefficiencies. Another extension would be to revisit the notion of "verifiability"by introducing third parties (e.g judges) who also acquire signals about preference characteristics and also update their beliefs from observing or interacting with the contracting parties.

Finally, we believe in the use of lab experiments to assess the importance of the effect of common p-belief perturbations on the likelihood that truth-telling will still occur in equilibrium. Preliminary

[20] Here ε-closeness means that at each state θ, f^ε selects the outcome $f(\theta)$ with probability $1-\varepsilon$.

[21] Note that in the Moore-Repullo setting (i.e. with quasi-linear utilities and arbitraily large transfers), for any social choice function f, we have the existence of a bad outcome (i.e. an outcome that, at each state of nature, is strictly worse for all players than any outcome in the range of the social choice function). In addition, because for each agent, there is no most preferred outcome, f also satisfies no-veto-power. Thus by Moore and Repullo (1990, Corollary 3, p.1094) f is Nash-implementable if and only if f is Maskin-monotonic.

work by Aghion, Fehr, Holden and Wilkening suggests that the effect is potentially large.[22]

References

[1] Abreu, D, and H. Matsushima, (1992) "Virtual Implementation in Iteratively Undominated Strategies: Complete Information", *Econometrica* 60, 993-1008

[2] Abreu, D. and A. Sen. (1991) "Virtual Implementation in Nash Equilibrium," *Econometrica*, 59, 997-1021.

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

[4] Aghion, P., E. Fehr, R. Holden and T. Wilkening. (2009), "Subgame Perfect Implementation: A Laboratory Experiment," unpublished working paper.

[5] Aghion P., D. Fudenberg, and R. Holden (2009) "Subgame Perfect Implementation With Almost Perfect Information," NBER working paper w15167.

[6] Che, Y. and D. Hausch (1999), "Cooperative Investments and the Value of Contracting," *American Economic Review* 89, 125-147.

[7] Cho, I.K. (1987), "A Refinement of Sequential Equilibrium," *Econometrica* 55, 1367-1389

[8] Cho, I.K. and D. Kreps (1987), "Signaling Games and Stable Equilibria", *Quarterly Journal of Economics,* 102, 179-221.

[9] Chung, K.S and J. Ely (2003), "Implementation with Near-Complete Information," *Econometrica* 71, 857-871.

[10] Cremer, J. and R.P. McLean (1988), "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions", *Econometrica* 56, 1247-1257.

[11] van Damme, E. and S. Hurkens (1997), "Games with Imperfectly Observable Commitment," *Games and Economic Behavior* 21, 282-308.

[12] Dekel, E. and D. Fudenberg (1990), "Rational Play Under Payoff Uncertainty," *Journal of Economic Theory* 52, 243-267.

[22] Aghion, Fehr, Holden and Wilkening (2009) conduct a laboratory experiment testing the robustness of a Moore-Repullo mechanism to common p-belief value perturbations. The experiment is meant to mimic the Hart-Moore example spelled out in Section 2. Subjects are randomly allocated to the buyer and seller roles, and play the mechanism ten times in a row. In one treatment there is complete information, in the other the subjects each receive a conditionally independent private signal which is 90% accurate-generated by the subjects drawing different colored balls from an urn. In the complete information treatment the proportion of buyers who announce low despite having a high signal declines from around 40% to 10% over the ten rounds. By contrast, in the incomplete information treatment buyers continue to lie more than 40% of the time. In periods 6-10 the average number of lies in the complete information treatment is 24%, whereas it is 42% in the incomplete information treatment.

[13] Duggan J. (1997) "Virtual Bayesian Implementation," *Econometrica* 65, 1175-1199.

[14] Fudenberg, D., D. Kreps, and D.K. Levine (1988), "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory* 44, 354-380.

[15] Fudenberg, D., D.K. Levine, and E. Maskin (1991), "Balanced-Budget Mechanisms for Adverse Selection Problems," unpublished working paper.

[16] Fudenberg D. and J. Tirole (1991a), *Game Theory*, MIT Press

[17] Fudenberg, D. and J. Tirole (1991b)"Perfect Bayesian and Sequential Equilibrium," *Journal of Economic Theory* 53 (1991), 236-260.

[18] Grossman, S, and O. Hart (1986), "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy* 94, 691-719.

[19] Hart, O. and J. Moore (2003), "Some (Crude) Foundations for Incomplete Contracts," unpublished working paper.

[20] Hendon, E., H.J. Jacobsen and B. Sloth, (1996) "The One-Shot Deviation Principle for Sequential Rationality," *Games and Economic Behavior* 12, 274-282.

[21] Jackson, M. (2001) "A Crash Course in Implementation Theory," *Social Choice and Welfare*, 18, 655-708.

[22] Johnson, S, J.W. Johnson and R.J. Zeckhauser (1990), "Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case," *Econometrica* 58, 873-900.

[23] Kreps, D.M. and R. Wilson (1982) "Sequential Equilibria," *Econometrica* 50, 863-894

[24] Kunimoto, T. (2010) "How Robust is Undominated Nash Implementation?," mimeo

[25] Kunimoto, T. and O. Tercieux (2009) "Implementation with Near-Complete Information: The Case of Subgame Perfection," mimeo

[26] Maskin, E. "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies* 66, 23-38.

[27] Maskin, E. and J. Moore (1999) "Implementation and Renegotiation," *Review of Economic Studies*, 66, 39-56.

[28] Maskin, E. and J. Tirole (1999a), "Unforeseen Contingencies and Incomplete Contracts," *Review of Economic Studies* 66, 83-114.

[29] Maskin, E. and J. Tirole (1999b), "Two Remarks on the Property-Rights Literature," *Review of Economic Studies* 66, 139-149.

[30] Matsushima, H. (1988) "A New Approach to the Implementation Problem," *Journal of Economic Theory*, 45, 128-144.

[31] Monderer, D. and D. Samet (1988), "Approximating Common Knowledge with Common Beliefs," *Games and Economic Behavior* 1, 170-190.

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

[33] Moore, J. and R. Repullo (1990), "Nash Implementation: A Full Characterization," *Econometrica*, 58, 1083-1099.

[34] Oury, M. and O. Tercieux (2009), "Continuous Implementation," working paper, available at http://www.pse.ens.fr/tercieux/ContUTP.pdf.

[35] Selten, R. (1965), "Spieltheoretische Behandlung Eines Oligopolmodells mit nachfragetragheit, " *Zeitschrift fur die gesamte Staatswissenschaft*, 121, 301-24.

6 Appendix

6.1 Proof of Theorem 2

Let a sequence of priors ν^{ε} (indexed by $\varepsilon > 0$) over the space of pair of signals and of states of nature be specified as in Section 3.2. By way of contradiction, assume that there exists a profile of mixed equilibrium strategies $\left\{\sigma^j_{k,l}, \mu^j_{k,l}, \rho^j_{k,l}, \tau^j_{k,l}\right\}$ such that for all $k \neq j$ and all l, $\sigma^j_{k,l}, \mu^j_{k,l}$ converges to 0 as $\varepsilon \to 0$; and for all $l \neq j$ and all k, $\rho^j_{k,l}$,$\tau^j_{k,l}$ converges to 0 as $\varepsilon \to 0$. We first need to show that the conditional probabilities of player 1 and 2 satisfy the following properties.

Lemma 1.
(i) For any announced ϕ_1 of player 2 at stage 2 (phase 1), any k and l, player 1's belief over Θ_1 converges to the belief given by his own signal:

$$Pr(\theta_1^k \mid s_1^{k,l}, \phi_1) \to 1 \ \ as\ \varepsilon \to 0.$$

(ii) Similarly, for any announced θ_1 of player 1 at stage 1 (phase 1), any k and l, player 2's belief over Θ_2 converges to the belief given by his own signal:

$$Pr(\theta_2^l \mid s_2^{k,l}, \theta_1) \to 1 \ \ as\ \varepsilon \to 0.$$

Proof. (i) Fix $\phi_1 = \theta_1^j$. We have[23] that for any $k, l \in \{1, ..., n\}$:

$$\begin{aligned}
&\Pr(\theta_1^k \mid s_1^{k,l}, \phi_1) \\
&\quad = \frac{\Pr(\theta_1^k, s_1^{k,l}, \phi_1)}{\Pr(\theta_1^k, s_1^{k,l}, \phi_1) + \sum_{\tilde{k} \neq k} \Pr(\theta_1^{\tilde{k}}, s_1^{k,l}, \phi_1)} \\
&\quad = \frac{1}{1 + \sum_{\tilde{k} \neq k} \Pr(\theta_1^{\tilde{k}}, s_1^{k,l}, \phi_1) \Big/ \Pr(\theta_1^k, s_1^{k,l}, \phi_1)}.
\end{aligned}$$

In the above expression, consider the following term:

$$\begin{aligned}
&\sum_{\tilde{k} \neq k} \Pr(\theta_1^{\tilde{k}}, s_1^{k,l}, \phi_1) \Big/ \Pr(\theta_1^k, s_1^{k,l}, \phi_1) \\
&= \frac{\sum_{\tilde{k} \neq k} \sum_{\tilde{l}} \sum_{(k_2,l_2)} \Pr(s_1^{k,l}, s_2^{k_2,l_2}, \theta_1^{\tilde{k}}, \theta_2^{\tilde{l}}) \mu^j_{k_2,l_2}}{\left\{ \begin{array}{c} \Pr(s_1^{k,l}, s_2^{k,l}, \theta_1^k, \theta_2^l) \mu^j_{k,l} + \sum_{(k_2,l_2) \neq (k,l)} \Pr(s_1^{k,l}, s_2^{k_2,l_2}, \theta_1^k, \theta_2^{l_2}) \mu^j_{k_2,l_2} \\ + \sum_{k_2} \sum_{\tilde{l}} \sum_{l_2 \neq \tilde{l}} \Pr(s_1^{k,l}, s_2^{k_2,l_2}, \theta_1^k, \theta_2^{\tilde{l}}) \mu^j_{k_2,l_2} \end{array} \right\}} \\
&= \frac{\sum_{\tilde{k} \neq k} \sum_{\tilde{l}} \sum_{(k_2,l_2)} \mu(\theta_1^{\tilde{k}}, \theta_2^{\tilde{l}}) \frac{\varepsilon^2}{n^4 - n^2} \mu^j_{k_2,l_2}}{\left\{ \begin{array}{c} \mu(\theta_1^k, \theta_2^l)[1 - \varepsilon - \varepsilon^2] \mu^j_{k,l} + \sum_{(k_2,l_2) \neq (k,l)} \mu(\theta_1^k, \theta_2^{l_2}) \frac{\varepsilon}{n^2 - 1} \mu^j_{k_2,l_2} \\ + \sum_{k_2} \sum_{\tilde{l}} \sum_{l_2 \neq \tilde{l}} \mu(\theta_1^k, \theta_2^{\tilde{l}}) \frac{\varepsilon^2}{n^4 - n^2} \mu^j_{k_2,l_2} \end{array} \right\}}
\end{aligned}$$

[23] It is easily shown that for $\varepsilon > 0$ small enough, for any $k, l \in \{1, ..., n\}$, we have $\Pr(\theta_1^k, s_1^{k,l}, \phi_1) > 0$.

If $\theta_1^j = \theta_1^k$, $\mu_{k,l}^j \to 1$ and so, in the denominator of the above expression, the term $\mu(\theta_1^k, \theta_2^l)[1-\varepsilon-\varepsilon^2]\mu_{k,l}^j$ tends to $\mu(\theta_1^k, \theta_2^l)$ while all the other terms tend to zero. So $\Pr(\theta_1^k \mid s_1^{k,l}, \phi_1)$ must tend to 1. If $\theta_1^j \neq \theta_1^k$, the term above can be rewritten as:

$$\frac{\sum_{\tilde{k}\neq k}\sum_{\tilde{l}}\sum_{(k_2,l_2)}\mu(\theta_1^{\tilde{k}},\theta_2^{\tilde{l}})\frac{1}{n^4-n^2}\mu_{k_2,l_2}^j}{\left\{\begin{array}{c}\mu(\theta_1^k,\theta_2^l)\frac{[1-\varepsilon-\varepsilon^2]}{\varepsilon^2}\mu_{k,l}^j+\sum_{(k_2,l_2)\neq(k,l)}\mu(\theta_1^k,\theta_2^{l_2})\frac{1}{n^2-1}\frac{1}{\varepsilon}\mu_{k_2,l_2}^j \\ +\sum_{k_2}\sum_{\tilde{l}}\sum_{l_2\neq\tilde{l}}\mu(\theta_1^k,\theta_2^{\tilde{l}})\frac{1}{n^4-n^2}\mu_{k_2,l_2}^j\end{array}\right\}} \tag{1}$$

We have that, by assumption, $\mu_{k_2,l_2}^j \to 1$ for $k_2 = j$, hence, the numerator in the above term tends to a strictly positive number while the terms on the denominator of the form $\mu(\theta_1^k, \theta_2^{l_2})\frac{1}{n^2-1}\frac{1}{\varepsilon}\mu_{j,l_2}^j$ must tend to $+\infty$ because $\mu_{k_2,l_2}^j \to 1$ for $k_2 = j$ and $j \neq k$. Hence the term in (1) tends to zero and so $\Pr(\theta_1^k \mid s_1^{k,l}, \phi_1) \to 1$. This completes the proof of (i). The proof of (ii) can be completed by mimicking the proof of (i). ■

Now we are in a position to complete the proof of Theorem 2.

Proof. We first note that for MR's logic to apply (i.e. for truthtelling to be the unique subgame perfect equilibrium under complete information), we must have[24] for each $\theta = (\theta_1, \theta_2)$:

$$u_1(f(\theta_1,\theta_2);\theta_1) > u_1(\{d;t_1\}(\theta_1);\theta_1) \tag{2}$$

and

$$u_2(f(\theta_1,\theta_2);\theta_2) > u_2(\{d;t_2\}(\theta_2);\theta_2) \tag{3}$$

and

$$u_2(f(\theta_1,\theta_2);\theta_2) > u_2(\{d;t_1\}(\theta_1);\theta_2) \tag{4}$$

and

$$u_1(f(\theta_1,\theta_2);\theta_1) > u_1(\{d;t_2\}(\theta_2);\theta_1). \tag{5}$$

Fix the prior ν^ε for $\varepsilon > 0$ small enough as defined in section 3.2., and consider the case where player 1 receives $s_1^{k,l}$. First by the above claim, for any message ϕ_1 of player 2 at stage 2, $\Pr(\theta_1^k \mid s_1^{k,l}, \phi_1) \to 1$ as $\varepsilon \to 0$. Hence, for $\varepsilon > 0$ small enough, player $s_1^{k,l}$ chooses $\{d;t_1\}(\theta_1^k)$ at stage 3 of phase 1 if he told the truth at stage 1 but player 2 challenged. Clearly, $\Pr(\theta_1^k, \theta_2^l, s_2^{k,l} \mid s_1^{k,l}) \to 1$ as $\varepsilon \to 0$. Hence, at stage 1, the expected payoff of player 1 for announcing θ_1^k converges toward $f(\theta_1^k, \theta_2^l)$ (recall that $\rho_{k,l}^k \to 1$ i.e. player 2 does not challenge with large probability) while if he lies at stage 1, his expected payoff converges toward $u_1(\{d;t_1\}(\theta_1^k), \theta_1^k)$ (recall that $\rho_{k,l}^k \to 1$ i.e. player 2 challenges with θ_1^k with a large probability). By Equation (2), there is no way that the equilibrium strategies $\left\{\sigma_{k,l}^j, \mu_{k,l}^j, \rho_{k,l}^j, \tau_{k,l}^j\right\}$ can make player 1 indifferent for $\varepsilon > 0$ small. Hence, player $s_1^{k,l}$ plays pure strategies in phase 1. Note now that player $s_1^{k,l}$ could deviate and claim that for $k' \neq k$, $\theta_1^{k'}$ is the true state. In this case, player 2 will believe with probability one that player 1 has received a

[24] By a slight abuse of notation, in the sequel, $u_i(\{d;t_j\}(\theta_j), \theta_i)$ denotes the utility of player i when j selects $\{d;t_j\}(\theta_j)$ at stage 3.

signal of the form $s_1^{k',\cdot}$. In addition, by the above lemma player 2 believes with high probability that $\theta = (\theta_1, \theta_2)$ where $\theta_2 = \theta_2^l$ is the true state. Since player 2 now believes with probability one that player 1 is of the form $s_1^{k',\cdot}$, she believes that player 1 will pick $\{d; t_1\}(\theta_1^{k'})$ at stage 3 in case she challenges him. Hence player $s_2^{k,l}$ will not challenge because she believes with high probability that $\theta = (\theta_1, \theta_2)$ where $\theta_2 = \theta_2^l$ is the true state and if she were to challenge, she expects player 1 to choose $\{d; t_1\}(\theta_1^{k'})$ at stage 3, while if she were to tell the truth, her expected payoff would tend to $u_2(f(\theta_1^{k'}, \theta_2^l), \theta_2^l)$. Hence by Equation (4), for $\varepsilon > 0$ small, player 2 will not challenge; thus, we get that $\mu_{k,l}^k = 0$ which is a contradiction. ■

6.2 Proof of Theorem 3

We first introduce some notations. Given a prior μ over $\Theta \times S$, we will sometimes abuse notation and write $\mu(\theta)$ for $[\text{marg}_\Theta \mu](\theta)$. Besides, given $s_{-i} \in S_{-i}$, we will also write $\mu(s_{-i})$ as $[\text{marg}_{S_{-i}} \mu](s_{-i})$. Finally, given some arbitrary countable space X, δ_x will denote the probability measure that puts probability 1 on $\{x\} \subset X$.

Let μ be any complete information prior, and assume that a mechanism Γ SPE-implements a non-Maskin monotonic SCF f. By hypothesis f is not Maskin monotonic, so there are θ' and θ'' satisfying (I) in the definition of Maskin monotonicity while $f(\theta') \neq f(\theta'')$. We now fix this particular θ' and θ'' throughout.

Since the mechanism Γ SPE-implements f, there exists a subgame perfect equilibrium $m_{\theta'}^*$ in $\Gamma(\theta')$ such that $g(m_{\theta'}^*) = f(\theta')$. Clearly, $m_{\theta'}^*$ is a Nash equilibrium of $\Gamma(\theta')$. From (I) in the definition of Maskin monotonicity, it follows that $m_{\theta'}^*$ is also a Nash equilibrium of $\Gamma(\theta'')$. Recall that $\mathcal{H}$ denotes the set of all possible histories. For each $t \geq 0$, let h_t^* be the history induced by $m_{\theta'}^*$ up to date t and let $\mathcal{H}^*$ denote the set of all such histories for any t. In addition, for each player i, let $\mathcal{H}_{-i}^*$ be the set of histories h along which every player $j \neq i$ has chosen the message $m_{\theta',j}^*(h)$; formally, $\mathcal{H}_{-i}^* \equiv \{h \in \mathcal{H} : h = (\emptyset, m^1, m^2, ..., m^{t-1})$ for some t and $m_j^{t'} = m_{j,\theta'}^{*,t'}$ for all $t' \leq t-1$ and all $j \neq i\}$. Note that $h_t^* \in \mathcal{H}_{-i}^*$ for each $t \geq 1$.

Consider the following family of information structures ν^ε. For each player i, let τ_i represent the profile of signals $s = (s_1, ..., s_n)$ defined by $s_i = s_i^{\theta''}$ and $s_j = s_j^{\theta'}$ for all $j \neq i$. For all i, ν^ε is given by[25]

$$\begin{aligned}
\nu^\varepsilon(\theta', \tau_i) &= \frac{\varepsilon}{n}\mu(\theta', s^{\theta'}); \\
\nu^\varepsilon(\theta', s^{\theta'}) &= (1-\varepsilon)\mu(\theta', s^{\theta'}); \text{ and} \\
\nu^\varepsilon(\tilde{\theta}, s^{\tilde{\theta}}) &= \mu(\tilde{\theta}, s^{\tilde{\theta}}) \ \ \forall \tilde{\theta} \neq \theta'.
\end{aligned}$$

In this information structure when the state is anything other than θ' or θ'', the state is common knowledge. Furthermore, when a player observes $s^{\theta'}$, he knows that the state is θ'. Obviously,

[25] This sequence of perturbations is similar to that used by Chung and Ely (2003). However, because sequential equilibrium requires verifying sequential rationality conditions that are not imposed by undominated Nash equilibrium, the body of proof is very different from that in Chung and Ely.

$\nu^\varepsilon \to \mu$ as $\varepsilon \to 0$. The support of ν^ε is denoted

$$\text{supp}(\nu^\varepsilon) = \{(\tilde{\theta}, s^{\tilde{\theta}}) : \tilde{\theta} \in \Theta\} \cup \{(\theta', \tau_i) : i \in N\}.$$

We build a sequential equilibrium $(\phi^\varepsilon, \sigma^\varepsilon)$ of $\Gamma(\nu^\varepsilon)$ where $g(\sigma^\varepsilon(s^{\theta''}); \emptyset) = f(\theta')$ for each $\varepsilon > 0$ small enough. This will show that there exist a sequence of priors $\{\nu^\varepsilon\}_{\varepsilon>0}$ that converges to μ and a corresponding sequence of sequential equilibria $\{(\phi^\varepsilon, \sigma^\varepsilon)\}_{\varepsilon>0}$ such that $g(\sigma^\varepsilon(s^{\theta''}); \emptyset) \to f(\theta') \neq f(\theta'')$ as ε goes to 0. This will complete the proof.

In the sequel, we will omit the dependence of σ^ε w.r.t. ε and simply write σ for σ^ε. In the following lines, we define a strategy σ and a family of system of beliefs Φ so that $g(\sigma(s^{\theta''}); \emptyset) = f(\theta')$. In addition, we will show that (ϕ, σ) is a sequential equilibrium of $\Gamma(\nu^\varepsilon)$ for some $\phi \in \Phi$. Φ and σ are defined as follows:

Definition of σ:

Σ1. For any player i and any $h_t \in \mathcal{H}^*$ or $h_t \notin \mathcal{H}^*_{-i}$, $\sigma_i(h_t, s_i^{\theta''}) = m^*_{i,\theta'}(h_t)$;[26]

Σ2. For any player i, any $h_t \in \mathcal{H}^*_{-i} \backslash \mathcal{H}^*$, $\sigma_i(h_t, s_i^{\theta''}) = \bar{m}_i(h_t)$ where $\bar{m}_i$ satisfies for any h_t,

$$h_t \in \mathcal{H}^* \text{ or } h_t \notin \mathcal{H}^*_{-i} \Rightarrow \bar{m}_i(h_t) = m^*_{i,\theta'}(h_t);$$
$$h_t \in \mathcal{H}^*_{-i} \backslash \mathcal{H}^* \Rightarrow \bar{m}_i(h_t) \in \arg\max \sum_{\tilde{\theta}} \nu^\varepsilon(\tilde{\theta}|s_i^{\theta''}) u_i(g(m'_i, m^*_{-i,\theta'}); h_t); \tilde{\theta})$$

where the max is taken over any m'_i that differs from $\bar{m}_i$ only at h.[27] By A1 there exists such $\bar{m}_i$**;**

Σ3. For any player i and any $h_t \in \mathcal{H}$, $\sigma_i(h_t, s_i^{\theta'}) = m^*_{i,\theta'}(h_t)$;

Σ4. And for any $h_t \in \mathcal{H}$, $\sigma_i(h_t, s_i^{\tilde{\theta}}) = m^*_{\tilde{\theta},i}(h_t)$ for $\tilde{\theta} \neq \theta', \theta''$ where $m^*_{\tilde{\theta}'}$ is an arbitrary subgame perfect equilibrium of $\Gamma(\tilde{\theta})$. (This is well-defined since f is implementable in subgame perfect equilibrium under complete information.)

Definition of Φ:

$\phi \in \Phi$ if and only ϕ satisfies the following three properties.

Φ1. Fix any $i \in N$, any $h_t \notin \mathcal{H}^*_{-i}$,

$$\phi_i \left[\cdot \middle| s_i^{\theta''}, h_t \right] = \delta_{(\theta', s^{\theta'}_{-i})}$$

[26] Note that players here send the messages that m prescribes for state θ' when their signal suggests the state is θ''.

[27] Note that the maximization above is over m'_i that differs from $\bar{m}_i$ only at h. Hence, since player i may be playing at several stages, it might be the case that this maximization depends on what player i is playing at further histories, and these further histories may be outside $\mathcal{H}^*_{-i} \backslash \mathcal{H}^*$ (for instance in case a player j different of i does not play according to $m^*_{j,\theta'}$ at some subsequent history). This is why we also have to define $\bar{m}_i$ outside $\mathcal{H}^*_{-i} \backslash \mathcal{H}^*$.

and

$$\text{supp}\left(\phi_i\left[\cdot|s_i^{\theta'}, h_t\right]\right) \subseteq \text{supp}\left(\nu^{\varepsilon}\left[\cdot|s_i^{\theta'}\right]\right)$$

and for all $l \neq i$ with $h_t \in \mathcal{H}^*_{-l} \setminus \mathcal{H}^*_{-i}$

(i.e., l has deviated from the path prescibed by $m^*_{\theta'}$)

$$\phi_i[(\theta', \tau_l) \mid s_i^{\theta'}, h_t] = 0.$$

Φ2. For any $i \in N$, any $h_t \in \mathcal{H}^*_{-i}$, any $s_i \in \{s_i^{\theta'}, s_i^{\theta''}\}$,

$$\phi_i[\cdot|s_i, h_t] = \nu^{\varepsilon}(\cdot|s_i).$$

Φ3. For any $i \in N$, any $h_t \in \mathcal{H}$ and any $s_i^{\tilde{\theta}} \notin \{s_i^{\theta'}, s_i^{\theta''}\}$, $\phi_i\left[\cdot|s_i^{\tilde{\theta}}, h_t\right] = \delta_{(\tilde{\theta}, s^{\tilde{\theta}}_{-i})}$ where δ_x denotes the probability measure that puts probability 1 on $\{x\}$.

Note that $h_T[\sigma(s^{\theta''}), \emptyset] = h_T[m^*_{\theta'}, \emptyset]$ and so, σ generates $g(\sigma(s^{\theta''}); \emptyset) = g(m^*_{\theta'}; \emptyset) = f(\theta')$. Hence, it only remains to show that (ϕ, σ) constitutes a sequential equilibrium for some $\phi \in \Phi$. In Section 6.2.1, we show that (ϕ, σ) satisfies sequential rationality for any $\phi \in \Phi$; and we establish that (ϕ, σ) satisfies consistency for some $\phi \in \Phi$ in Section 6.2.2.

6.2.1 Sequential rationality

Fix any $\phi \in \Phi$. Sequential rationality of (ϕ, σ) will be proved by Claims 1 and 2.

Claim 1. *For any $i \in N$, $s_i \neq s_i^{\theta''}, h_t \in \mathcal{H}$:*

$$\sum_{(\tilde{\theta}, s_{-i})} \phi_i[(\tilde{\theta}, s_{-i})|s_i, h_t]\left[u_i(g(\sigma(s); h_t); \tilde{\theta}) - u_i(g(\sigma_i'(s_i), \sigma_{-i}(s_{-i}); h_t); \tilde{\theta})\right] \geq 0$$

for each σ_i'.

Claim 1 states that for any player i with any signal $s_i \neq s_i^{\theta''}$, σ_i is a best response to σ_{-i} given his belief ϕ_i. This will be checked by considering three classes of histories: (1) Histories where all players have played according to the equilibrium $m^*_{\theta'}$ (i.e. in $\mathcal{H}^*$); (2) histories where player i has not played according to $m^*_{i,\theta'}$ but all other players have (i.e. in $\mathcal{H}^*_{-i} \setminus \mathcal{H}^*$) and finally (3) histories where some player other than i has not played according to $m^*_{\theta'}$ (i.e. outside $\mathcal{H}^*_{-i}$).

In particular, in the non-trivial case where $s_i = s_i^{\theta'}$, we will show that for any of these histories h_t, whenever player i follows σ_i against σ_{-i}, player i believes with probability one that the outcome will be given by $g(m^*_{\theta'}; h_t)$, while if player i deviates from $\sigma_i(s_i)$ to some m_i' , player i believes with probability one that the outcome will be given by $g(m_i', m^*_{-i,\theta'}; h_t)$. Because $m^*_{\theta'}$ is a subgame perfect equilibrium in the complete information game $\Gamma(\theta')$ and player i with signal $s_i^{\theta'}$ believes with probability one that θ' is the true state, this will prove the claim.

Proof of Claim 1. Fix any player i. Claim 1 is obvious for $s_i^{\tilde{\theta}} \neq s_i^{\theta'}$ because by $\mathbf{\Phi 3}$, $\phi_i\left[\cdot\middle|s_i^{\tilde{\theta}}, h_t\right] = \delta_{(\tilde{\theta}, s_{-i}^{\tilde{\theta}})}$ and so state $\tilde{\theta}$ is common knowledge. By $\mathbf{\Sigma 4}$, we can further conclude that $\sigma(s^{\tilde{\theta}}) = m_{\tilde{\theta}}^*$ is a subgame perfect equilibrium in the complete information game $\Gamma(\tilde{\theta})$. Hence, we focus on the case where $s_i = s_i^{\theta'}$. By construction, $\nu^{\varepsilon}(\theta'|s_i^{\theta'}) = 1$ and so this player knows the state is θ', and he knows the profile of signals is either $s^{\theta'}$ or τ_k for some $k \neq i$. We partition the set of all histories into three classes $\mathcal{H}^*$; $\mathcal{H}_{-i}^* \backslash \mathcal{H}^*$ and $\mathcal{H} \backslash \mathcal{H}_{-i}^*$ and consider the following three cases: Case (1) $h_t \in \mathcal{H}^*$; Case (2) $h_t \in \mathcal{H}_{-i}^* \backslash \mathcal{H}^*$; and Case (3) $h_t \notin \mathcal{H}_{-i}^*$.

- Case (1): $h_t \in \mathcal{H}^*$

  In this case, each player has played according to $m_{\theta'}^*$ and if players $j \neq i$ received signals of either $s_j^{\theta'}$ or $s_j^{\theta''}$, by $\mathbf{\Sigma 1}$ and $\mathbf{\Sigma 3}$, this will continue to be the case as long as all players conform to σ. So when players are playing strategy σ, and the profile of signals received is $s^{\theta'}$ or τ_k**,** for $k \neq i$ any subsequent history also falls into $\mathcal{H}^*$. Thus, $g(\sigma(s^{\theta'}); h_t) = g(\sigma(\tau_k); h_t) = g(m_{\theta'}^*; h_t)$.

  Now suppose player i deviates to a strategy σ_i' so that $\sigma_i'(s_i^{\theta'}) = m_i'$. Clearly, since $m_i' \neq \sigma_i(s_i^{\theta'})$, there is a date at which player i does not play according to $m_{i,\theta'}^*$. Thus, by $\mathbf{\Sigma 1}$ and $\mathbf{\Sigma 3}$, when the profile of signals received is either $s^{\theta'}$ or τ_k for $k \neq i$, any subsequent history of h_t either falls in $\mathcal{H}^*$ (player i has played according to $m_{i,\theta'}^*$ so far) or does not fall in $\mathcal{H}_{-k}^*$ for each $k \neq i$ (at some point in this history, player i has not played according to $m_{i,\theta'}^*$). In each of these cases, again by $\mathbf{\Sigma 1}$ and $\mathbf{\Sigma 3}$, player i's opponents are playing according to $m_{-i,\theta'}^*$. So we get [28]

  $$g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(s_{-i}^{\theta'}); h_t) = g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(\tau_k); h_t) = g(m_i', m_{-i,\theta'}^*; h_t).$$

  Here again, since $m_{\theta'}^*$ is a subgame perfect equilibrium in the complete information game $\Gamma(\theta')$, we have

  $$u_i(g(m_{\theta'}^*; h_t); \theta') \geq u_i(g(m_i', m_{-i,\theta'}^*; h_t); \theta').$$

  Thus, we get $u_i(g(\sigma(s^{\theta'}); h_t); \theta') \geq u_i(g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(s_{-i}^{\theta'}); h_t); \theta')$ and $u_i(g(\sigma(\tau_k); h_t); \theta') \geq u_i(g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(\tau_k); h_t); \theta')$ for each $k \neq i$. Now since by $\mathbf{\Phi 2}$, $\phi_i[\cdot \mid s_i^{\theta'}, h_t]$ may assign strictly positive weight only to $(\theta', s_{-i}^{\theta'})$ and (θ', τ_k) for each $k \neq i$, we can conclude

  $$\sum_{(\tilde{\theta}, s_{-i})} \phi_i[(\tilde{\theta}, s_{-i})|s_i^{\theta'}, h_t] \left[u_i(g(\sigma_i(s_i^{\theta'}), \sigma_{-i}(s_{-i}); h_t); \tilde{\theta}) - u_i(g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(s_{-i}); h_t); \tilde{\theta})\right] \geq 0.$$

- Case (2): $h_t \in \mathcal{H}_{-i}^* \backslash \mathcal{H}^*$

  Since $h_t \in \mathcal{H}_{-i}^*$ and $h_t \notin \mathcal{H}^*$, only player i has not played according to $m_{i,\theta'}^*$**.** Then, it is clear that h_t does not fall in $\mathcal{H}_{-k}^*$ for each $k \neq i$ (recall that $\mathcal{H}_{-k}^*$ is the set of histories under which every player j other than k has played according to $m_{j,\theta'}^*$**).** It is also clear that any subsequent history does not fall in $\mathcal{H}_{-k}^*$ for each $k \neq i$**.** By $\mathbf{\Sigma 1}$ and $\mathbf{\Sigma 3}$, we thus obtain

[28] We abuse notation because we should use $\sigma_{-i}(\tau_l \backslash s_i^{\theta'})$ instead of $\sigma_{-i}(\tau_l)$.

that each player k other than i will play according to $m^*_{k,\theta'}$ at any subsequent history when receiving signal $s_k^{\theta'}$ or $s_k^{\theta''}$. Hence

$$g(\sigma(s^{\theta'}); h_t) = g(\sigma(\tau_k); h_t) = g(m^*_{\theta'}; h_t).$$

Consider the case where player i deviates to a strategy σ_i' so that $\sigma_i'(s_i^{\theta'}) = m_i'$. Here, since (by a similar argument as above) any history that player i can achieve by deviating does not fall in $\mathcal{H}^*_{-k}$ for each $k \neq i$, each player k other than i will be playing according to $m^*_{k,\theta'}$ at any subsequent history whether he receive $s_k^{\theta'}$ or $s_k^{\theta''}$ which implies

$$g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(s_{-i}^{\theta'}); h_t) = g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(\tau_k); h_t) = g(m_i', m^*_{-i,\theta'}; h_t).$$

Since $m^*_{\theta'}$ is a subgame perfect equilibrium in the complete information game $\Gamma(\theta')$, we already have $u_i(g(m^*_{\theta'}; h_t); \theta') \geq u_i(g(m_i', m^*_{-i,\theta'}; h_t); \theta')$. Thus, we also get

$$\begin{aligned} &u_i(g(\sigma(s^{\theta'}); h_t); \theta') \geq u_i(g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(s_{-i}^{\theta'}); h_t); \theta') \text{ and} \\ &u_i(g(\sigma(\tau_k); h_t); \theta') \geq u_i(g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(\tau_k); h_t); \theta') \text{ for each } k \neq i. \end{aligned}$$

Now, since by **Φ2** we know that $\phi_i[\cdot|s_i^{\theta'}, h_t]$ assigns a strictly positive weight only to $(\theta', s_{-i}^{\theta'})$ and (θ', τ_k) for each $k \neq i$, we can conclude

$$\sum_{(\tilde{\theta}, s_{-i})} \phi_i[(\tilde{\theta}, s_{-i})|s_i^{\theta'}, h_t] \left[u_i(g(\sigma(s_i^{\theta'}), \sigma_{-i}(s_{-i}), h_t); \tilde{\theta}) - u_i(g(\sigma_i'(s_i^{\theta'}), \sigma_{-i}(s_{-i}); h_t); \tilde{\theta})\right] \geq 0.$$

- Case (3): $h_t \notin \mathcal{H}^*_{-i}$

  In this case, at least one player $j \neq i$ has not played according to $m^*_{j,\theta'}$.

  By **Σ3**, we know that when each player j receives signal $s_j^{\theta'}$ then these players play according to $m^*_{j,\theta'}$, so $\sigma(s^{\theta'}) = m^*_{\theta'}$. Thus, at history h_t, the oucome achieved by playing σ when the profile of signals is $s^{\theta'}$ must be the same as the one when playing $m^*_{\theta'}$, i.e.

  $$g(\sigma(s^{\theta'}); h_t) = g(m^*_{\theta'}; h_t).$$

  In addition, for each $l \neq i$ with $h_t \notin \mathcal{H}^*_{-l}$, by definition, some player j other than l has not played according to $m^*_{j,\theta'}$ and obviously this will continue to be the case at any subsequent histories. Hence, any subsequent histories does not belong to $\mathcal{H}^*_{-l}$ either. At any such histories, we know by **Σ1**, that player l will be playing according to $m^*_{l,\theta'}$ when he receives $s_l^{\theta''}$ while when players j other than l receive signal $s_j^{\theta'}$ by **Σ3** they will also be playing according to $m^*_{j,\theta'}$. Hence, we get that the outcome achieved from history h_t when playing σ and when the profile of signals received is τ_l is equal to the outcome achieved from history h_t when

playing $m^*_{\theta'}$. Otherwise stated, for each $l \neq i$ with $h_t \notin \mathcal{H}^*_{-l}$, we have

$$g(\sigma(\tau_l); h_t) = g(m^*_{\theta'}; h_t).$$

Now, when player i deviates say to a strategy σ'_i so that $\sigma'_i(s_i^{\theta'}) = m'_i$, using the argument above, when the other players receive signal profile $s_{-i}^{\theta'}$, we know that the outcome achieved is

$$g(\sigma'_i(s_i^{\theta'}), \sigma_{-i}(s_{-i}^{\theta'}); h_t) = g(m'_i, m^*_{-i,\theta'}; h_t).$$

while for each $l \neq i$ with $h_t \notin \mathcal{H}^*_{-l}$, we know that

$$g(\sigma'_i(s_i^{\theta'}), \sigma_{-i}(\tau_l); h_t) = g(m'_i, m^*_{-i,\theta'}; h_t).$$

Since $m^*_{\theta'}$ is a subgame perfect equilibrium in the complete information game $\Gamma(\theta')$, we have $u_i(g(m^*_{\theta'}; h_t); \theta') \geq u_i(g(m'_i, m^*_{-i,\theta'}; h_t); \theta')$. Thus, we get

$$u_i(g(\sigma(s^{\theta'}); h_t); \theta') \geq u_i(g(\sigma'_i(s_i^{\theta'}), \sigma_{-i}(s_{-i}^{\theta'}); h_t); \theta')$$

and for each $l \neq i$ such that $h_t \notin \mathcal{H}^*_{-l}$, $u_i(g(\sigma(\tau_l); h_t); \theta') \geq u_i(g(\sigma'_i(s_i^{\theta'}), \sigma_{-i}(\tau_l); h_t); \theta')$. Because by **Φ1**, $\phi_i[\cdot|s_i^{\theta'}, h_t]$ may assign strictly positive weight only to $(\theta', s_{-i}^{\theta'})$ and (θ', τ_l) for each $l \neq i$ such that $h_t \notin \mathcal{H}^*_{-l}$, we can conclude

$$\sum_{(\tilde{\theta}, s_{-i})} \phi_i[(\tilde{\theta}, s_{-i})|s_i^{\theta'}, h_t] \left[u_i(g(\sigma(s_i^{\theta'}, s_{-i}); h_t); \tilde{\theta}) - u_i(g(\sigma'_i(s_i^{\theta'}), \sigma_{-i}(s_{-i}); h_t); \tilde{\theta})\right] \geq 0.$$

This completes the proof.

■

Claim 2. *For any $i \in N$, $s_i = s_i^{\theta''}$, and $h_t \in \mathcal{H}$:*

$$\sum_{(\tilde{\theta}, s_{-i})} \phi_i[(\tilde{\theta}, s_{-i})|s_i, h_t] \left[u_i(g(\sigma(s); h_t); \tilde{\theta}) - u_i(g((\sigma'_i(s_i), \sigma_{-i}(s_{-i})); h_t); \tilde{\theta})\right] \geq 0$$

for each σ'_i.

Claim 2 states that for any player i with any signal $s_i^{\theta''}$, σ_i is a best response to σ_{-i} given his belief ϕ_i. Here again we consider the same partition of histories as in Claim 1. When h_t is a history where each player has played according to $m^*_{\theta'}$ (i.e. $h_t \in \mathcal{H}^*$), player i assigns positive probability to both θ'' and θ'. However, we will show that here again player i believes with probability one that the other players will be playing according to $m^*_{-i,\theta'}$, whether he deviates or not. Hence, if he does not deviate and $h_t \in \mathcal{H}^*$, he gets $f(\theta')$ while if he deviates to m'_i he gets $g(m'_i, m^*_{-i,\theta'}; h_t)$. Because $m^*_{\theta'}$ is a subgame perfect equilibrium in $\Gamma(\theta')$, we know that the deviation is not profitable if θ' is

the true state, and monotonicity (condition (I)) implies that this is also not profitable if the state is θ''. Since these are the only states to which player i assigns strictly positive probability, this will complete the argument for this class of histories.

The easy case occurs when h_t is a history where a player other than i has not played according to $m^*_{\theta'}$ (i.e. $h_t \notin \mathcal{H}^*_{-i}$). In such a case, player i believes with probability one that θ' is the true state. In addition we will check that whenever player i uses σ_i against σ_{-i}, player i believes with probability one that the outcome will be given by $g(m^*_{\theta'}; h_t)$, while if player i deviates from $\sigma_i(s_i)$ to m'_i, player i believes with probability one that the outcome will be given by $g(m'_i, m^*_{-i,\theta'}; h_t)$. Here again, the fact that $m^*_{\theta'}$ is a subgame perfect equilibrium in the complete information game will lead to the desired result. Finally, in the last case where player i has not played according to $m^*_{\theta'}$ while all other players have (i.e. $h_t \in \mathcal{H}^*_{-i} \backslash \mathcal{H}^*$), we will also check that player i assigns probability one to his opponent playing $m^*_{-i,\theta'}$. But σ_i has been constructed (see **Σ2**) so that playing σ_i is better than any one-shot deviation. Then the one-shot deviation principle for sequential equilibrium will complete the proof of Claim 2. Taken together, Claims 1 and 2 establish sequential rationality of (ϕ, σ).

Proof of Claim 2. This claim will be proved by studying three different cases depending on the type of history we consider: (1) $h_t \in \mathcal{H}^*$; (2) $h_t \notin \mathcal{H}^*_{-i}$ and (3) $h_t \in \mathcal{H}^*_{-i} \backslash \mathcal{H}^*$.

- Case (1): $h_t \in \mathcal{H}^*$

  In this case, each player has played according to $m^*_{\theta'}$**.** Note that, by **Σ1** and **Σ3**, if each player j received signals of either $s_j^{\theta'}$ or $s_j^{\theta''}$, by **Σ1** and **Σ3**, this will continue to be the case as long as all players conform to σ. So when players are playing strategy σ, and player i's opponents received either signal profile $s_{-i}^{\theta'}$ or $s_{-i}^{\theta''}$**,** any subsequent history also falls into $\mathcal{H}^*$. Thus,

  $$g(\sigma(s_i^{\theta''}, s_{-i}^{\theta''}); h_t) = g(\sigma(s_i^{\theta''}, s_{-i}^{\theta'}); h_t) = g(m^*_{\theta'}; h_t).$$

  Now suppose that player i deviates to a strategy σ'_i so that $\sigma'_i(s_i^{\theta''}) = m'_i$. Since $m'_i \neq \sigma_i(s_i^{\theta''})$, there must exist a date at which player i does not play according to $m^*_{i,\theta'}$. Thus, by **Σ1** and **Σ3**, when player i's opponents receive signal $s_{-i}^{\theta'}$ or $s_{-i}^{\theta''}$, any subsequent history of h_t either falls in $\mathcal{H}^*$ (player i has played according to $m^*_{i,\theta'}$ so far) or does not fall in $\mathcal{H}^*_{-k}$ for each $k \neq i$ (at some point in this history, player i has not played according to $m^*_{i,\theta'}$). In each of these cases, by **Σ1** and **Σ3**, player i's opponents are playing according to $m^*_{-i,\theta'}$. So we get

  $$g(\sigma'_i(s_i^{\theta''}), \sigma_{-i}(s_{-i}^{\theta'}); h_t) = g(\sigma'_i(s_i^{\theta''}), \sigma_{-i}(s_{-i}^{\theta''}); h_t) = g(m'_i, m^*_{-i,\theta'}; h_t). \tag{6}$$

  Here again, since $m^*_{\theta'}$ is a subgame perfect equilibrium in the complete information game $\Gamma(\theta')$, we have

  $$u_i(g(m^*_{\theta'}; h_t); \theta') \geq u_i(g(m'_i, m^*_{-i,\theta'}; h_t); \theta').$$

Thus, we also get

$$u_i(g(\sigma(s_i^{\theta''}, s_{-i}^{\theta'}); h_t); \theta') \geq u_i(g(\sigma_i'(s_i^{\theta''}), \sigma_{-i}(s_{-i}^{\theta'}); h_t); \theta'). \tag{7}$$

The above inequality together with (6) also implies

$$u_i(g(\sigma(s_i^{\theta''}, s_{-i}^{\theta''}); h_t); \theta') \geq u_i(g(\sigma_i'(s_i^{\theta''}), \sigma_{-i}(s_{-i}^{\theta''}); h_t); \theta').$$

Since $g(\sigma(s_i^{\theta''}, s_{-i}^{\theta''}); h_t) = g(m_{\theta'}^*; h_t^*) = f(\theta')$ and we have assumed that θ' and θ'' are two states satisfying (I) in the definition of Maskin monotonicity, we get that

$$u_i(g(\sigma(s_i^{\theta''}, s_{-i}^{\theta''}); h_t); \theta'') \geq u_i(g(\sigma_i'(s_i^{\theta''}), \sigma_{-i}(s_{-i}^{\theta''}); h_t); \theta''). \tag{8}$$

Now since by **Φ2**, $\phi_i[\cdot|s_i^{\theta''}, h_t]$ assigns a strictly positive weight only to $(\theta', s_{-i}^{\theta'})$ and $(\theta'', s_{-i}^{\theta''})$, we conclude with (7) and (8):

$$\sum_{(\tilde{\theta}, s_{-i}))} \phi_i[(\tilde{\theta}, s_{-i}) \mid s_i^{\theta''}, h_t] \left[u_i(g(\sigma(s_i^{\theta''}, s_{-i}); h_t); \tilde{\theta}) - u_i(g(\sigma_i'(s_i^{\theta''}), \sigma_{-i}(s_{-i}); h_t); \tilde{\theta})\right] \geq 0.$$

- Case (2): $h_t \notin \mathcal{H}_{-i}^*$

  In this case, at least one player $j \neq i$ has not played according to $m_{j,\theta'}^*$; this is still the case for any subsequent histories, so that they all fall outside $\mathcal{H}_{-i}^*$. By **Σ1**, if player i plays according to σ_i, from h_t, he will play according to $m_{i,\theta'}^*$. Now, by **Σ3**, we know that when player j other than i receives signal $s_j^{\theta'}$, then he plays according to $m_{j,\theta'}^*$. Thus, the outcome achieved when the profile of signals is $(s_i^{\theta''}, s_{-i}^{\theta'})$ must be the same as the outcome achieved when $m_{\theta'}^*$ is played i.e.we obtain

  $$g(\sigma(s_i^{\theta''}, s_{-i}^{\theta'}); h_t) = g(m_{\theta'}^*; h_t).$$

  Suppose player i deviates to a strategy σ_i' so that $\sigma_i'(s_i^{\theta''}) = m_i'$. Since if the other players are receiving signal profile $s_{-i}^{\theta'}$, they will all be playing according to $m_{-i,\theta'}^*$, we obtain

  $$g(\sigma_i'(s_i^{\theta''}), \sigma_{-i}(s_{-i}^{\theta'}); h_t) = g(m_i', m_{-i,\theta'}^*; h_t).$$

  Since $m_{\theta'}^*$ is a subgame perfect equilibrium in the complete information game $\Gamma(\theta')$, we have $u_i(g(m_{\theta'}^*; h_t); \theta') \geq u_i(g(m_i', m_{-i,\theta'}^*; h_t); \theta')$. Thus, we also get

  $$u_i(g(\sigma(s_i^{\theta''}, s_{-i}^{\theta'}); h_t); \theta') \geq u_i(g(\sigma_i'(s_i^{\theta''}), \sigma_{-i}(s_{-i}^{\theta'}); h_t); \theta').$$

  Because by **Φ1**, $\phi_i[(\theta', s_{-i}^{\theta'}) \mid s_i^{\theta''}, h_t] = 1$, so we can conclude

  $$\sum_{(\tilde{\theta}, s_{-i})} \phi_i[(\tilde{\theta}, s_{-i})|s_i^{\theta''}, h_t] \left[u_i(g(\sigma(s_i^{\theta''}, s_{-i}); h_t); \tilde{\theta}) - u_i(g(\sigma_i'(s_i^{\theta''}), \sigma_{-i}(s_{-i}); h_t); \tilde{\theta})\right] \geq 0.$$

- Case (3): $h_t \in \mathcal{H}^*_{-i} \backslash \mathcal{H}^*$

  Since $h_t \in \mathcal{H}^*_{-i}$ and $h_t \notin \mathcal{H}^*$, only player i has not played according to $m^*_{i,\theta'}$. Then h_t does not fall in $\mathcal{H}^*_{-k}$ for each $k \neq i$ (recall that $\mathcal{H}^*_{-k}$ is the set of histories under which every player j other than k has played according to $m^*_{j,\theta'}$**).** It is also clear that any subsequent history does not fall in $\mathcal{H}^*_{-k}$ for each $k \neq i$**.** By **Σ1** and **Σ3**, whether player i's opponents have received $s^{\theta'}_{-i}$ or $s^{\theta''}_{-i}$, they all play according to $m^*_{-i,\theta'}$. By **Φ2** we know that $\phi_i[\cdot|s^{\theta''}_i, h_t] = \nu^\varepsilon(\cdot|s^{\theta''}_i)$ assigns a strictly positive weight only to $(\theta', s^{\theta'}_{-i})$ and $(\theta'', s^{\theta''}_{-i})$. In addition, we have that for any $h \in \mathcal{H}^*$ or $h \notin \mathcal{H}^*_{-i} : \sigma_i(h, s^{\theta''}_i) = m^*_{i,\theta'}(h, s^{\theta''}_i)$. Since $h_t \in \mathcal{H}^*_{-i} \backslash \mathcal{H}^*$, we conclude with **Σ2**

$$\sum_{(\tilde{\theta}, s_{-i})} \nu^\varepsilon(\tilde{\theta}, s_{-i}|s^{\theta''}_i) \left[u_i(g(\sigma(s^{\theta''}_i, s_{-i}); h_t); \tilde{\theta}) - u_i(g(\sigma'_i(s^{\theta''}_i), \sigma_{-i}(s_{-i}); h_t); \tilde{\theta}) \right] \geq 0$$

  for any σ'_i that differs from σ_i only at h_t. By this and case (1) and (2), we know that at any history players have no profitable one-shot deviation, by the one-shot deviation principle (see Hendon, Jacobsen, and Sloth (1996)[29]) this yields:

$$\sum_{(\tilde{\theta}, s_{-i})} \nu^\varepsilon(\tilde{\theta}, s_{-i}|s^{\theta''}_i) \left[u_i(g(\sigma(s^{\theta''}_i, s_{-i}); h_t); \tilde{\theta}) - u_i(g(\sigma'_i(s^{\theta''}_i), \sigma_{-i}(s_{-i}); h_t); \tilde{\theta}) \right] \geq 0$$

  for any σ'_i. This completes the proof.

■

6.2.2 Consistency

In this section, we show that for some $\phi \in \Phi$, (ϕ, σ) satisfies consistency.

To show this part, we first fix σ as defined above and consider the following sequence $\{(\phi^k, \sigma^k)\}_{k=0}^\infty$ of assessments. Let $\eta_k > 0$ for each k and $\eta_k \to 0$ as $k \to \infty$. For each player i, $h_t \in \mathcal{H}$, and signal s_i, let $\xi_i(h_t, s_i, \cdot)$ be any strictly positive prior over $M_i(h_t) \backslash \{\sigma_i(s_i, h_t)\}$ and define σ^k_i as

$$\sigma^k_i(m^t_i \mid h_t, s^{\theta''}_i) = \begin{cases} 1 - \eta^{T \times n}_k \text{ if } m^t_i = \sigma_i(h_t, s^{\theta''}_i); \\ \eta^{T \times n}_k \times \xi_i(h_t, s^{\theta''}_i, m^t_i) \text{ otherwise} \end{cases}$$

where T is the (finite) length of the longest final history; and for any signal $s_i \neq s^{\theta''}_i$:

$$\sigma^k_i(m^t_i \mid h_t, s_i) = \begin{cases} 1 - \eta_k \text{ if } m^t_i = \sigma_i(h_t, s_i); \\ \eta_k \times \xi_i(h_t, s_i, m^t_i) \text{ otherwise} \end{cases}.$$

Let ϕ^k be the unique consistent belief associated with each σ^k. It is easy to check that σ^k converges

[29] Hendon, Jacobsen, and Sloth (1996) assume that for each i and h, $M_i(h)$ is finite, which is our A1. It is easy to check that their argument goes through in case $M_i(h)$ is countably infinite.

to σ and also that ϕ^k converges[30]. Let $\phi \equiv \lim_{k\to\infty} \phi^k$. In what follows, we show that ϕ satisfies **Φ1, Φ2** and **Φ3**. This will show that (ϕ, σ) satisfies consistency, and $\phi \in \Phi$ as claimed.

To do so, we explicitly compute each ϕ^k and study its limit as k tends to infinity. In general for each $(\tilde{\theta}, \tilde{s}_{-i}) \in \Theta \times S_{-i}$, each $h_t = (m^1, ..., m^{t-1}) \in \mathcal{H}$, and each $\tilde{s}_i \in S_i$, we have

$$\phi_i^k[(\tilde{\theta}, \tilde{s}_{-i}) \mid \tilde{s}_i, h_t] = \frac{\nu^\varepsilon(\tilde{\theta}, \tilde{s}_{-i}, \tilde{s}_i) \times \prod_{t'=1}^{t-1} \left[\sigma^k(m^{t'} \mid h_{t'}, \tilde{s})\right]}{\sum\limits_{(\hat{\theta}, s'_{-i})} \nu^\varepsilon(\hat{\theta}, s'_{-i}, \tilde{s}_i) \times \prod_{t'=1}^{t-1} \left[\sigma^k(m^{t'} \mid h_{t'}, s'_{-i}, \tilde{s}_i)\right]}.$$

In the above formula for each $t' \leq t$, $h_{t'}$ stands for the truncation of h_t to the first t' elements i.e., $h_{t'} = (m^1, ..., m^{t'-1})$.

Claim 3. *ϕ satisfies* **Φ1**.

Claim 3 says that for any player i who sees signal $s_i^{\theta''}$ and has an opportunity to play after some other player has not played according to $m^*_{\theta'}$ (i.e. $h_t \notin \mathcal{H}^*_{-i}$), then under $\phi \equiv \lim_{k\to\infty} \phi^k$, player i believes with probability one that the state is θ', and that the other players have received $s_{-i}^{\theta'}$. In order to show that, we observe that if every player other than i has received a signal $s_j \in \{s_j^{\theta'}, s_j^{\theta''}\}$, then at such a history some player j other than i has deviated from σ. Then since under the sequence of totally mixed strategies built above, it is (infinitely) more likely (as η_k tends to 0) that a deviation occured at $s_j^{\theta'}$ rather than at $s_j^{\theta''}$, in the limit, Bayes rule will then put probability one on $s_j^{\theta'}$ and given that the prior ν^ε assigns strictly positive weight only to $(\theta'', s_{-i}^{\theta''})$ and $(\theta', s_{-i}^{\theta'})$, Bayes rule will then put probability arbitrarily close to one on $(\theta', s_{-i}^{\theta'})$. In case, player i received the private signal $s_i^{\theta'}$, if h_t is a history under which all players other than l have played according to $m^*_{\theta'}$ (i.e. $h_t \in \mathcal{H}^*_{-l}$), then the deviating player is l and again using a similar argument as above, we show that player i must assign probability 0 to l receiving $s_l^{\theta''}$ and so to τ_l**.**

Consider player i, $h_t \notin \mathcal{H}^*_{-i}$. The proof is reduced to checking the following two cases:

Proof of Claim 3. **Case 1:** $s_i = s_i^{\theta''}$

[30] As will become clear from the proof, the sequence $\{\phi^k\}_k$ does converge. Moreover, convergence in the definition of consistency is taken uniformly over messages and histories. In the case where $M_i(h)$ is countably infinite (we will discuss this case in Section 6.3 in the Appendix), two natural convergence notions can be used: *point-wise* convergence or *uniform* convergence. The set of sequential equilibria is smaller when one assumes uniform convergence. Hence, the use of uniform convergence strengthens our result.

Recall that $\nu^\varepsilon(\cdot, s_i^{\theta''})$ assigns a strictly positive weight only to $(\theta'', s_{-i}^{\theta''})$ and $(\theta', s_{-i}^{\theta'})$. Hence,

$$\begin{aligned}
&\phi_i^k[(\theta', s_{-i}^{\theta'}) \mid s_i^{\theta''}, h_t] \\
&= \frac{\nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta''}) \times \prod_{j\neq i}\prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'})}{\nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta''}) \times \prod_{j\neq i}\prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'}) + \nu^\varepsilon(\theta'', s_{-i}^{\theta''}, s_i^{\theta''}) \times \prod_{j\neq i}\prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta''})} \\
&= \frac{\nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta''})}{\nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta''}) + \nu^\varepsilon(\theta'', s_{-i}^{\theta''}, s_i^{\theta''}) \times \frac{\prod_{j\neq i}\prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} | h_{t'}, s_j^{\theta''})}{\prod_{j\neq i}\prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} | h_{t'}, s_j^{\theta'})}}.
\end{aligned}$$

We now show that the ratio $\prod_{j\neq i}[\prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta''}) \Big/ \prod_{j\neq i}\prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'})]$ tends to 0 as k tends to infinity. This will show that $\phi_i^k[(\theta', s_{-i}^{\theta'}) \mid s_i^{\theta''}, h_t] \to 1$ and $\phi_i^k[(\theta'', s_{-i}^{\theta''}) \mid s_i^{\theta''}, h_t] \to 0$.

Note first that in case every player j other than i receive signal $s_j \in \{s_j^{\theta'}, s_j^{\theta''}\}$, there must exist a player $\hat{\jmath} \neq i$ and a date $\hat{t} \leq t-1$ so that $\hat{\jmath}$ has not played according to $\sigma_{\hat{\jmath}}$, i.e. $\sigma_{\hat{\jmath}}(h_{\hat{t}}, s_{\hat{\jmath}}) \neq m_{\hat{\jmath}}^{\hat{t}}$. To see this, proceed by contradiction and assume that $\sigma_{-i}(h_{t'}, s_{-i}) = m_{-i}^{t'}$ for all $t' \leq t-1$. This implies that whenever $h_{t'-1} \in \mathcal{H}_{-i}^*$, we must have $h_{t'} \in \mathcal{H}_{-i}^*$, because $h_{t'-1} \in \mathcal{H}_{-i}^*$ implies that either $h_{t'-1} \in \mathcal{H}^*$ (i.e., no player has deviated) or $h_{t'-1} \notin \mathcal{H}_{-j}^*$ for all $j \neq i$ (i.e., i has deviated). In either case, $\sigma_{-i}(h_{t'-1}, s_{-i}) = m_{-i,\theta'}^*(h_{t'-1})$ is obtained by **Σ1** and **Σ3**. Since we have assumed that $\sigma_{-i}(h_{t'-1}, s_{-i}) = m_{-i}^{t'-1}$, we get $m_{-i}^{t'-1} = m_{-i,\theta'}^*(h_{t'-1})$, which proves that $h_{t'} \in \mathcal{H}_{-i}^*$. Since $h_1 = \emptyset \in \mathcal{H}^* \subseteq \mathcal{H}_{-i}^*$, this simple inductive argument shows that $h_t \in \mathcal{H}_{-i}^*$, a contradiction.

By construction of σ^k, this implies that for some $\hat{\jmath} \neq i$ and $\hat{t} \leq t-1$:

$$\sigma_{\hat{\jmath}}^k(m_{\hat{\jmath}}^{\hat{t}} \mid h_{\hat{t}}, s_{\hat{\jmath}}^{\theta''}) = \eta_k^{T\times n} \xi_{\hat{\jmath}}(h_{\hat{t}}, s_{\hat{\jmath}}^{\theta''}, m_{\hat{\jmath}}^{\hat{t}}). \tag{9}$$

Now, we have:

$$\begin{aligned}
&\frac{\prod_{j\neq i}\prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta''})}{\prod_{j\neq i}\prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'})} \leq \frac{\eta_k^{T\times n} \times \xi_{\hat{\jmath}}(h_{\hat{t}}, s_{\hat{\jmath}}^{\theta''}, m_{\hat{\jmath}}^{\hat{t}}) \times 1}{\prod_{j\neq i}\prod_{t'=1}^{t-1} \eta_k \xi_j(h_{t'}, s_j^{\theta'}, m_j^{t'})} \\
&= \frac{\eta_k^{T\times n}}{\eta_k^{(t-1)(n-1)}} \times \frac{\xi_{\hat{\jmath}}(h_{\hat{t}}, s_{\hat{\jmath}}^{\theta''}, m_{\hat{\jmath}}^{\hat{t}})}{\prod_{j\neq i}\prod_{t'=1}^{t-1} \xi_j(h_{t'}, s_j^{\theta'}, m_j^{t'})} \to 0 \ \text{ (as } k \to \infty).
\end{aligned}$$

Here, the inequality is assured by (9) and the construction of σ^k that, for all j and $t' \leq t-1$, $\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'}) \geq \eta_k \times \xi_j(h_{t'}, s_j^{\theta'}, m_j^{t'})$.

Case 2: $s_i = s_i^{\theta'}$

Recall that $\nu^\varepsilon(\cdot, s_i^{\theta'})$ assigns a strictly positive weight only to $(\theta', s_{-i}^{\theta'})$ and (θ', τ_l) for each $l \neq i$.

Hence,

$$\phi_i^k[(\theta',\tau_l) \mid s_i^{\theta'}, h_t]$$

$$= \frac{\nu^\varepsilon(\theta',\tau_l) \times \prod_{j\neq l,i} \prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'}) \times \prod_{t'=1}^{t-1} \sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta''})}{\sum_{z\neq i} \nu^\varepsilon(\theta',\tau_z) \times \prod_{j\neq z,i} \prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'}) + \nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta'}) \times \prod_{j\neq i} \prod_{t'=1}^{t-1} \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'})}$$

$$= \frac{\nu^\varepsilon(\theta',\tau_l)}{\sum_{z\neq i} \nu^\varepsilon(\theta',\tau_z) \times c_z(k) + \nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta'}) \times \prod_{t'=1}^{t-1} \sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta'}) \Big/ \prod_{t'=1}^{t-1} \sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta''})}$$

for some positive functions $c_z(k)$. We now show that if $h_t \in \mathcal{H}^*_{-l}$, then the ratio

$$\prod_{t'=1}^{t-1} \sigma_l^k(m_l^{t'}, h_{t'}, s_l^{\theta'}) \Big/ \prod_{t'=1}^{t-1} \sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta''})$$

tends to ∞ as k tends to infinity. This will show that $\phi_i^k[(\theta',\tau_l) \mid s_i^{\theta'}, h_t] \to 0$ for all l such that $h_t \in \mathcal{H}^*_{-l}$; and hence that ϕ satisfies **Φ1.** Assume that $h_t \in H^*_{-l}$ for some l, as we already claimed, if all players j other than i have received a signal $s_j \in \{s_j^{\theta'}, s_j^{\theta''}\}$, there is a player $\hat{\jmath} \neq i$ and a date $\hat{t} \leq t-1$ so that $\hat{\jmath}$ has not played according to $\sigma_{\hat{\jmath}}$ i.e. $\sigma_{\hat{\jmath}}(h_{\hat{t}}, s_{\hat{\jmath}}) \neq m_{\hat{\jmath}}^{\hat{t}}$**.** Now, since $h_t \in \mathcal{H}^*_{-l}$, we claim that $\hat{\jmath} = l$. Indeed, $h_t \in \mathcal{H}^*_{-l}$ means that any player j other than l has played according to $m^*_{j,\theta'}$. So if player l had played according to σ_l (i.e. for all $t' : \sigma_l(h_{t'}, s_l) = m_l^{t'}$), repeated applications of **Σ1** and **Σ3** would yield to $h_t = h_t^* \in \mathcal{H}^*_{-i}$ which is false by assumption.

By construction of σ^k, this implies that there exists $\hat{t} \leq t-1$ such that $\sigma_l(h_{\hat{t}}, s_l) \neq m_l^{\hat{t}}$ and so:

$$\sigma_l^k(m_l^{\hat{t}} \mid h_{\hat{t}}, s_l^{\theta''}) = \eta_k^{T\times n} \xi_l(h_{\hat{t}}, s_l^{\theta''}, m_l^{\hat{t}}). \tag{10}$$

Now, we have

$$\frac{\prod_{t'=1}^{t-1} \sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta'})}{\prod_{t'=1}^{t-1} \sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta''})} \geq \frac{\eta_k^{t-1} \prod_{t'=1}^{t-1} \xi_l(h_{t'}, s_l^{\theta'}, m_l^{t'})}{\eta_k^{T\times n} \xi_l(h_{\hat{t}}, s_l^{\theta''}, m_l^{\hat{t}}) \times 1} \to \infty \text{ (as } k \to \infty).$$

Where the inequality is assured by (10) and (assuming wlog that η_k is small) we use the fact that by construction, for all $t' \leq t-1$, $\sigma_l^k(m_j^{t'} \mid h_{t'}, s_l^{\theta'}) \geq \eta_k \times \xi_l(h_{t'}, s_l^{\theta'}, m_l^{t'})$. ■

Claim 4. *ϕ satisfies* **Φ2.**

Claim 4 says that if a player i gets signal $s_i^{\theta'}$ or $s_i^{\theta''}$ then at a history h_t under which each of

his opponent have played according to $m^*_{\theta'}$; ϕ is the same as his beliefs given only by his private signal.

To prove this, we show that if every player other than player i has received a signal $s_j \in \{s_j^{\theta'}, s_j^{\theta''}\}$ then at histories where each player other than i has played according to $m^*_{\theta'}$, each player other than i has played according to σ at each previous stage. This ensures that for any $h_t \in \mathcal{H}^*_{-i}$, no player other than i has deviated from the candidate sequential equilibrium strategy σ and so i's beliefs must be given by his private signal**.**

Proof of Claim 4. Consider player i, $h_t \in \mathcal{H}^*_{-i}$. Here again, the proof is reduced to checking the following two cases.

Case 1: $s_i = s_i^{\theta''}$

Recall that $\nu^\varepsilon(\cdot, s_i^{\theta''})$ assigns a strictly positive weight only to $(\theta'', s_{-i}^{\theta''})$ and $(\theta', s_{-i}^{\theta'})$. Hence,

$$\begin{aligned}
&\phi_i^k[(\theta'', s_{-i}^{\theta''}) \mid s_i^{\theta''}, h_t] \\
&= \frac{\nu^\varepsilon(\theta'', s_{-i}^{\theta''}, s_i^{\theta''}) \times \prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta''})}{\nu^\varepsilon(\theta'', s^{\theta''}) \times \prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta''}) + \nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta''}) \times \prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'})} \\
&= \frac{\nu^\varepsilon(\theta'', s_{-i}^{\theta''}, s_i^{\theta''})}{\nu^\varepsilon(\theta'', s^{\theta''}) + \nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta''}) \times \frac{\prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'}|h_{t'}, s_j^{\theta'})}{\prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'}|h_{t'}, s_j^{\theta''})}}
\end{aligned}$$

We now show that the ratio

$$\prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'}) \Big/ \prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta''}) \to 1 \quad \text{as } k \to \infty.$$

This will show that $\phi_i^k[(\theta'', s_{-i}^{\theta''}) \mid s_i^{\theta''}, h_t] \to \nu^\varepsilon((\theta'', s_{-i}^{\theta''}) \mid s_i^{\theta''})$ and $\phi_i^k[(\theta', s_{-i}^{\theta'}) \mid s_i^{\theta''}, h_t] \to \nu^\varepsilon((\theta', s_{-i}^{\theta'}) \mid s_i^{\theta''})$**.**

Note now that if players $j \neq i$ receive signal $s_j \in \{s_j^{\theta'}, s_j^{\theta''}\}$, then for all $t' \leq t-1$, $\sigma_j(h_{t'}, s_j) = m_j^{t'}$. To see this, note that for any $t' \leq t-1 : h_{t'} \in \mathcal{H}^*_{-i}$, thus, either every player has played according to $m^*_{\theta'}$ (i.e. $h_{t'} \in \mathcal{H}^*$) or player i has not played according to $m^*_{i,\theta'}$ (i.e. $h_{t'} \notin \mathcal{H}^*_{-j}$ for all $j \neq i$). In each of these cases we know, by **Σ1** and **Σ3**, that σ_j presribes to play according to $m^*_{j,\theta'}$. Since $h_{t'} \in \mathcal{H}^*_{-i}$ this implies that $\sigma_j(h_{t'}, s_j) = m^*_{j,\theta'}(h_{t'}) = m_j^{t'}$.

By construction of σ^k, this in turn implies that for all $j \neq i$ and $t' \leq t-1$:

$$\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'}) = 1 - \eta_k \text{ and } \sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta''}) = 1 - \eta_k^{T\times n}$$

Thus,

$$\prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'}) \Big/ \prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta''}) \to 1 \text{ as } k \to \infty$$

Case 2: $s_i = s_i^{\theta'}$

Recall that $\nu^\varepsilon(\cdot, s_i^{\theta'})$ assigns a strictly positive weight only to $(\theta', s_{-i}^{\theta'})$ and (θ', τ_l) for $l \neq i$. Hence,

$$\phi_i^k[(\theta', s_{-i}^{\theta'}) \mid s_i^{\theta'}, h_t]$$

$$= \frac{\nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta'}) \times \prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'})}{\nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta'}) \times \prod_{j\neq i}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'}) + \sum_{l\neq i}\nu^\varepsilon(\theta', \tau_l) \times \prod_{j\neq i,l}\prod_{t'=1}^{t-1}\sigma_j^k(m_j^{t'} \mid h_{t'}, s_j^{\theta'}) \times \prod_{t'=1}^{t-1}\sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta''})}$$

$$= \frac{\nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta'})}{\nu^\varepsilon(\theta', s_{-i}^{\theta'}, s_i^{\theta'}) + \sum_{l\neq i}\nu^\varepsilon(\theta', \tau_l) \times \frac{\prod_{t'=1}^{t-1}\sigma_l^k(m_l^{t'}|h_{t'}, s_l^{\theta''})}{\prod_{t'=1}^{t-1}\sigma_l^k(m_l^{t'}|h_{t'}, s_l^{\theta'})}}$$

We now show that for each $l \neq i$, the ratio $\prod_{t'=1}^{t-1}[\sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta''}) \Big/ \prod_{t'=1}^{t-1}[\sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta'})$ tends to 1 as k tends to infinity. This will show that $\phi_i^k[(\theta', s_{-i}^{\theta'}) \mid s_i^{\theta'}, h_t] \to \nu^\varepsilon((\theta', s_{-i}^{\theta'}) \mid s_i^{\theta'})$ and similar reasoning shows that for each $l \neq i : \phi_i^k[(\theta', \tau_l) \mid s_i^{\theta'}, h_t] \to \nu^\varepsilon((\theta', \tau_l) \mid s_i^{\theta'})$; and hence, ϕ satisfies **Φ2**.

Now, by a similar reasoning as in the case above, we get that for all $l \neq i$ and $t' \leq t-1$:

$$\sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta'}) = 1 - \eta_k \text{ and } \sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta''}) = 1 - \eta_k^{T\times n}$$

Thus,

$$\prod_{t'=1}^{t-1}\sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta''}) \Big/ \prod_{t'=1}^{t-1}\sigma_l^k(m_l^{t'} \mid h_{t'}, s_l^{\theta'}) \to 1 \text{ as } k \to \infty$$

■

Finally, observing that for $s_i^{\tilde{\theta}} \notin \{s_i^{\theta'}, s_i^{\theta''}\}$, $\nu^\varepsilon(\cdot, s_i^{\tilde{\theta}})$ assigns a weight one to $(\check{\theta}, s_{-i}^{\tilde{\theta}})$, we have established the following claim, which completes the proof of Theorem 3.

Claim 5. *ϕ satisfies* **Φ3.**

6.3 Theorem 3 extends to countable messages

Here we extend Theorem 3 to mechanisms that have countably infinite message spaces. This extension is important because the literature on full implementation theory often uses integer games where each player has to announce an integer and becomes the dictator when his integer is the largest one, as in Maskin (1999) and in Moore and Repullo (1988).

Assumption A2. *$M_i(h)$ is countable for each i and h.*

The next assumption says that against any profile of strategy in the complete information game, in the neighborhood of complete information, each player i has a non-empty set of best responses. This condition is vacuously satisfied under A1, so Theorems 3 and 4 show that if a mechanism can implement a non-monotonic social choice function both under complete information and under small common p-belief value perturbations, then under this mechanism agents must have not well-defined best responses. In addition, we show in the supplemental materials that when the state space is finite Moore and Repullo's general mechanism has well-defined best-responses (under weak assumptions) and so our argument also applies there.

Assumption A3. *The sequential mechanism Γ has well-defined best replies: for any player i, any $\theta \in \Theta$, any $m_{-i} \in M_{-i}$, there exists $\bar{\xi}(i,\theta,m_{-i}) > 0$ such that for any $\beta \in \Delta(\Theta)$ with $\beta(\theta) \geq 1 - \bar{\xi}(i,\theta,m_{-i})$, for any $m_i \in M_i$ we have for all $h \in \mathcal{H}$:*

$$\arg\max \sum_{\tilde{\theta}} \beta(\tilde{\theta}) u_i(g\left(\left(m_i', m_{-i}\right)\right); h); \tilde{\theta}) \neq \emptyset$$

where the max *is taken over any m_i' that differs from m_i only at h.*

Remark 4. *If the mechanism is not finite but the set of outcomes is, A3 is also vacuously satisfied. We also note that A3 is not needed for sequential mechanisms in which each player moves only once.*[31]

Theorem 4. *Assume A2 and A3. Suppose that a mechanism Γ SPE-implements a non-monotonic SCF f. Fix any complete information prior μ. There exist a sequence of priors $\{\nu^\varepsilon\}_{\varepsilon>0}$ that converges to μ and a corresponding sequence of sequential equilibria $\{(\phi^\varepsilon, \sigma^\varepsilon)\}_{\varepsilon>0}$ such that as ε tends to 0, $g(\sigma^\varepsilon(s^\theta); \emptyset) \nrightarrow f(\theta)$ for some $\theta \in \Theta$.*

Proof. The proof is essentially the same as the proof of Theorem 3 where we only consider finite mechanisms. So, we claim that there are essentially only two changes we need to extend the proof of Theorem 3 to the case of countably infinite message spaces. First, in the beginning of the proof of Theorem 3, we have to choose $\varepsilon > 0$ small enough to apply A3. Second, we will show that A3 guarantees that **Σ2** (which is introduced in the proof of Theorem 3) is well defined. This will be proved in the next subsection. ■

6.3.1 Additional material: A3 guarantees that Σ2 is well-defined

Fix $\varepsilon > 0$ small enough so that $\nu^\varepsilon(\theta' | s_i^{\theta'}) \geq 1 - \bar{\xi}(i,\theta,m^*_{-i,\theta})$. We shall claim that A3 guarantees that one can construct $\bar{m}_i$ needed for **Σ2**. First, for any $h_t \in \mathcal{H}^*$ or $h_t \notin \mathcal{H}^*_{-i}$, we set $\bar{m}_i(h_t) =$

[31] One can directly check this in the definition of strategy σ (**Σ2**) used in the proof of Theorem 3. More specifically, it can be checked there that for each player, A3 is only used at histories where this player has to choose a message and at which he has previously deviated from the equilibrium. By construction, there is no such history.

$m^*_{i,\theta}(h_t)$. Second, we define $\bar{m}_i$ by induction on the set of histories in $\mathcal{H}^*_{-i}\backslash\mathcal{H}^*$. Take any history $h_t \in \mathcal{H}^*_{-i}\backslash\mathcal{H}^*$ so that there is no subsequent history that falls into $\mathcal{H}^*_{-i}\backslash\mathcal{H}^*$. Since we already defined $\bar{m}_i(h_t) = m^*_{i,\theta}(h_t)$ for any $h_t \notin \mathcal{H}^*_{-i}\backslash\mathcal{H}^*$, $\bar{m}_i$ has been defined for any subsequent histories. By A3 we obtain

$$\arg\max \sum_{\tilde{\theta}} \nu^{\varepsilon}(\tilde{\theta}|s_i^{\theta'}) u_i(g\left(\left(m'_i, m^*_{-i,\theta}\right)\right); h_t); \tilde{\theta}) \neq \emptyset$$

where the max is taken over any m'_i that differs from $\bar{m}_i$ only at h_t and are identical at any subsequent histories (what happens before h_t is obviously irrelevant).

Now set

$$\bar{m}_i(h_t) \in \arg\max \sum_{\tilde{\theta}} \nu^{\varepsilon}(\tilde{\theta}|s_i^{\theta'}) u_i(g\left(\left(m'_i, m_{-i}\right)\right); h_t); \tilde{\theta}).$$

This establishes that one can inductively construct $\bar{m}_i$ so that $\bar{m}_i$ satisfies the properties needed for **Σ2**.

6.3.2 Additional material: A3 is satisfied in the Moore-Repullo canonical mechanism

We will review some of the main results of Moore and Repullo (1988) here.

Definition 2 (Moore and Repullo (1988))**.** *A social choice correspondence f satisfies Condition C if, for every pair of profiles $\theta, \phi \in \Theta$ with $a \in f(\theta)\backslash f(\phi)$, there exists a finite sequence*

$$\sigma(\theta, \phi; a) \equiv \{a_0 = a, a_1, \ldots, a_k, \ldots, a_l, a_{l+1}\} \subset A,$$

with $l = l(\theta, \phi; a) \geq 1$, such that:

1. *for each $k = 0, \ldots, l-1$, there is some particular player $j(k) = j(k|\theta, \phi; a)$, for whom*

$$u_{j(k)}(a_k; \theta) \geq u_{j(k)}(a_{k+1}; \theta);$$

2. *there is some player $j(l) = j(l|\theta, \phi; a)$ for whom*

$$u_{j(l)}(a_l; \theta) \geq u_{j(l)}(a_{l+1}; \theta) \text{ and } u_{j(l)}(a_{l+1}; \phi) > u_{j(l)}(a_l; \phi).$$

Further, $l(\theta, \phi; a)$ is uniformly bounded by some $\bar{l} < \infty$.

Assuming Condition C holds, let $\mathcal{Q}(f)$ be a class of subsets Q of A. A typical Q is defined as follows:

> For each pair of profiles θ and ϕ in Θ, and for each $a \in f(\theta)\backslash f(\phi)$, select *one* sequence $\sigma(\theta, \phi; a)$ satisfying (1) and (2) in Condition C. Then let Q be the union of the elements in these sequences.

$\mathcal{Q}(f)$ comprises the Q's constructed from all possible selections.

Definition 3. *A social choice correspondence f satisfies Condition C^+ if it satisfies Condition C and the following condition as well: there exists a particular $Q^+ \in \mathcal{Q}(f)$, and a particular set $B \subset A$ containing Q^+, such that the following is true for each $\theta \in \Theta$:*

- *Each player i has nonempty maximal set $B_i^*(\theta) \subset B$ under θ i.e. $B_i^*(\theta) = \arg\max_{a \in B} u_i(a; \theta)$.*
- *$B_i^*(\theta) \cap B_j^*(\theta) = \emptyset$ for each $\theta \in \Theta$ and each $i, j \in N$ with $i \neq j$*
- *$B_i^*(\theta) \cap Q^+ = \emptyset$ for each i and each θ.*

Let the selected sequences $\sigma(\theta, \phi; a) \in Q^+$ be labelled $\sigma^+(\theta, \phi; a)$. We restrict our attention to social choice functions. Define the Moore-Repullo canonical mechanism $\Gamma^{MR} = (M, g)$ as follows.

Stage 0: each player i announces some triplet $m_{i,0} = (\theta^i, a^i, n_0^i)$, where $\theta^i \in \Theta, a^i = f(\theta^i)$, and n_0^i is a nonnegative integer. There are three possibilities to consider:

1. all n players agree on a common profile θ and outcome $a = f(\theta)$, then outcome a is chosen. STOP
2. If only $n-1$ players agree on a common profile θ and outcome $a \in f(\theta)$, and if the remaining player i announces a profile ϕ, and
   (a) if $a = f(\phi)$, then outcome a is implemented; STOP
   (b) if $a \neq f(\phi)$ but i is not the agent $j(0)$ prescribed in $\sigma^+(\theta, \phi; a)$, then outcome a is implemented; STOP
   (c) if $a \neq f(\phi)$ and $i = j(0)$, then go to Stage 1.
3. If neither (1) nor (2) apply, then the player with the highest integer n_0^i is allowed to choose an outcome from B. Ties are broken by selecting from the players who announced the highest number according to who has the smallest i. STOP

Stage $k = 1, \ldots, l$: each player i can either raise a "flag," or announce a nonnegative integer $n_k^i \in \mathbb{N}$, i.e., $m_{i,k} \in M_{i,k} \in \{\text{flag}\} \cup \mathbb{N}$. Again there are three possibilities to consider:

1. If $n-1$ or more flags are raised, then the agent $j(k-1)$ prescribed in $\sigma^+(\theta, \phi; a)$ is allowed to choose an outcome from B. STOP
2. If $n-1$ or more players announce zero, and
   (a) if the player $j(k)$ prescribed in $\sigma^+(\theta, \phi; a)$ is one of those who announce zero, then implement outcome a_k from sequence $\sigma^+(\theta, \phi; a)$; STOP
   (b) if $j(k)$ does not announce zero, then
      i. if $k < l$, go to Stage $k+1$;

ii. if $k = l$, implement outcome a_{l+1} from sequence $\sigma^+(\theta, \phi; a)$. STOP

(c) If neither (1) nor (2) apply, then the player who announces the highest integer n_k^i is allowed to choose an outcome from B. STOP

Theorem 5 (Moore and Repullo (1988))**.** *If a social choice function f satisfies Condition C^+, and $n \geq 3$, then f can be implemented in subgame perfect equilibrium.*

Moore and Repullo (1988) show the above theorem by using the mechanism described above. We note that this mechanism satisfies A3 if the set of outcomes A is finite or when each player's preferences over A are strict and utilities are bounded. Furthermore, the above mechanism satisfies A3 whenever (i) the B given in condition C^+ is a compact set of outcomes; (ii) $u_i : A \times \Theta \to \mathbb{R}$ is continuous in a.[32,33] It is worth noting that mechanisms that check C^+ and then appeal to Moore and Repullo (1988)'s result often assume (i) and (ii). This is the case for instance in Moore and Repullo (1988)'s examples of risk-sharing (Section 6.1) or the production contract example (Section 6.2). More importantly, it is also the case in Maskin and Tirole (1999a)'s proof of the irrelevance Theorem. Hence our non-robusntess result (Theorem 4) also apply to Maskin and Tirole's irrelevance Theorem.

[32] Then, for any $\beta \in \Delta(\Theta)$, $\arg\max_{a \in B} \sum_{\tilde{\theta}} \beta(\tilde{\theta}) u_i(a; \tilde{\theta}) \neq \emptyset$. We note that a one-shot deviation of player i at stage k in Γ^{MR} allows player i possibly to fall into an integer game at stage k where he can get any outcome in B; if he cannot fall into this integer game, he can only induce a finite number of outcomes, say B_k, by deviating. In any case, he has a most preferred deviation i.e. $\arg\max_{a \in B} \sum_{\tilde{\theta}} \beta(\tilde{\theta}) u_i(a; \tilde{\theta}) \neq \emptyset$, $\arg\max_{a \in B \cup B_k} \sum_{\tilde{\theta}} \beta(\tilde{\theta}) u_i(a; \tilde{\theta}) \neq \emptyset$, and $\arg\max_{a \in B_k} \sum_{\tilde{\theta}} \beta(\tilde{\theta}) u_i(a; \tilde{\theta}) \neq \emptyset$. Then A3 is satisfied whenever (i) and (ii) hold.

[33] Note that A2 need not be satisfied for these mechanisms since B need not be countable. A2 was introduced only to define sequential equilibrium in a simple manner. If one uses perfect Bayesian equilibrium instead, we believe that A2 is not required.