SUBGAME-PERFECT IMPLEMENTATION UNDER INFORMATION PERTURBATIONS*

PHILIPPE AGHION
DREW FUDENBERG
RICHARD HOLDEN
TAKASHI KUNIMOTO
OLIVIER TERCIEUX

We consider the robustness of extensive form mechanisms to deviations from common knowledge about the state of nature, which we refer to as *information perturbations*. First, we show that even under arbitrarily small information perturbations the Moore-Repullo mechanism does not yield (even approximately) truthful revelation and that in addition the mechanism has sequential equilibria with undesirable outcomes. More generally, we prove that any extensive form mechanism is fragile in the sense that if a non-Maskin monotonic social objective can be implemented with this mechanism, then there are arbitrarily small information perturbations under which an undesirable sequential equilibrium also exists. Finally, we argue that outside options can help improve efficiency in asymmetric information environments, and that these options can be thought of as reflecting ownership of an asset. *JEL* Codes: C72, D23, D78, D82.

I. 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, that

*This article builds on two preliminary contributions, respectively, by Aghion, Fudenberg, and Holden (2009) and Kunimoto and Tercieux (2009). 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, Eric Maskin, Roger Myerson, Antonio Penta, Andrew Postlewaite, Jean Tirole, Jorgen Weibull, Ivan Werning, Tom Wilkening, Muhamet Yildiz, 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, the Canadian Institute for Advanced Research, and the referees and editor of this journal for very useful comments and suggestions. Thanks also to Ashley Cheng for careful proofreading. Financial support from Canadian Institute of Advanced Research (CIFAR) (Aghion), from National Science grants SES 0648616, 0954162 (Fudenberg), and from Fonds Québécois de la Recherche sur la Société et la Culture (FQRSC), Social Sciences and Humanities Research Council (SSHRC) of Canada, Japan Society for the Promotion of Science (JSPS), and the Seimeikai Foundation (Kunimoto) is gratefully acknowledged.

The Quarterly Journal of Economics (2012), 1843–1881. doi:10.1093/qje/qjs026.
Advance Access publication on August 30, 2012.

is, which *social choice rules*, can be implemented by mechanisms that respect the players' incentives. Although only *Maskin 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.[1]

This article considers the robustness of subgame-perfect implementation to arbitrarily small amounts of incomplete information about the state of nature θ, which we refer to as "information perturbations."[2] It is known that refinements of Nash equilibrium are not robust to general small perturbations of the payoff structure (Fudenberg, Kreps, and Levine 1988, henceforth FKL), but our results do not follow from theirs as we consider a more restrictive class of perturbations: we fix the map from states to payoffs and perturb the prior distribution over the states of the world and signal structure, so in particular the 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, henceforth MR) result which roughly says that for any social choice rule, 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 (which we call an MR mechanism) inducing truth-telling as the unique subgame-perfect equilibrium. As in MR, our focus is on *exact* implementation, where "exact implementation" means that we require the set of equilibrium outcomes to *exactly* coincide with those picked by the rule.[3]

1. Recall that a social choice rule or function f is *Maskin monotonic* if for any pair of states θ and θ' such that $a = f(\theta)$, and a never goes down in the preference ranking of any agent when moving from state θ to state θ', then necessarily $a = f(\theta')$.

2. It follows from Theorem 14.5 of Fudenberg and Tirole (1991a: 567) that under our small informational perturbations, for each profile of signals that has strictly positive probability under complete information, some state of nature is *common p-belief* (Monderer and Samet 1989) with p arbitrarily close to 1. That is, everybody believes this is the true state with probability at least p, everybody believes with probability at least p that everybody believes this is the true state with probability at least p and so on.

3. Much of the implementation literature studies exact implementation. Virtual implementation (Matsushima 1988; Abreu and Sen 1991) uses nondeterministic mechanisms, and only requires that social choice rules be implemented

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 equilibria whose outcomes differ from those of the rule.

Our first result shows that MR mechanisms can only robustly satisfy the first requirement of exact implementation if the rule that is implemented is Maskin monotonic. That is, whenever an MR mechanism implements a non-Maskin monotonic social choice rule, the truth-telling equilibrium ceases to be an equilibrium in some nearby environment. More specifically, we show that an MR mechanism that implements a social choice rule f under common knowledge (or complete information[4]) about the state of nature does not yield even approximately truthful revelation under arbitrarily small information perturbations, if this f is not Maskin monotonic.[5]

We then move beyond MR mechanisms to consider *any* extensive-form mechanism. Our second result is concerned with the nonrobustness of the second requirement of exact implementation: namely, whenever any mechanism implements a non-Maskin monotonic social choice rule, 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 spaces,[6] then given any mechanism that "subgame-perfect" implements a non-Maskin monotonic social choice rule 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 information perturbations (i.e., of deviations from complete information about the state of nature) and a corresponding sequence of sequential equilibria

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. In addition, virtual implementation provides incentives for renegotiation on the equilibrium path: as Abreu and Matsuhima (1992) acknowledge, virtual implementation supposes that the social planner can commit ex ante to outcomes that will be known at the time of implementation to be highly inefficient.

4. Throughout the article, we use "complete information" and "common knowledge" interchangeably.

5. As we shall stress in Section II.E below, Maskin monotonicity is precisely the property that the social choice rules usually considered in contract theory do not satisfy.

6. The Online Appendix extends the result to the case of countable message spaces.

for the mechanism under the corresponding information perturbations, whose outcomes do not converge to $f(\theta)$ for at least one state θ. In other words, there always exist arbitrarily small information perturbations under which an "undesirable" sequential[7] (and hence perfect Bayesian) equilibrium exists.

Three 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 arbitrarily small information perturbations 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-Maskin monotonic social choice rule under common knowledge has at least one Nash equilibrium which is not a subgame-perfect equilibrium; we prove that this undesirable Nash equilibrium can be turned into an undesirable sequential equilibrium by only introducing small information perturbations. The third insight is that there is a role for asset ownership to mitigate the investment and trade inefficiencies that arise when the contracting parties have private information ex post about the state of nature θ.

Our results are not a straightforward application of those on the robustness of refinements of Nash equilibrium because we consider a smaller class of perturbations. While FKL consider several nested classes of perturbations, even the most restrictive form they analyze allows a player's payoff in the perturbed game to vary with the realized actions in an arbitrary way. In the mechanism design setting, this implies that some (low-probability) "crazy types" might have a systematic preference for truth telling. Because the messages and outcome functions of the mechanism are not primitives but endogenous objects to be chosen by the social planner, it may seem natural to restrict the perturbations to be independent of the messages and depend only on the allocation that is implemented.

Our article contributes most directly to the mechanism design literature, starting with Maskin's (1999) Nash implementation

7. We remind the reader of the formal definition in Section IV.B.

result and Moore and Repullo's (1988) subgame-perfect implementation analysis, by showing the nonrobustness of subgame-perfect implementation to information perturbations.[8] Our article is also related to Chung and Ely's (2003) study of the robustness of undominated Nash implementation. Chung and Ely show that if a social choice rule is not Maskin monotonic but can be implemented in undominated Nash equilibrium[9] under complete information, then there are information perturbations under which an undesirable undominated Nash equilibrium appears. In contrast, we consider extensive-form mechanisms and show that only Maskin 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.[10] Hence, although our article has a similar spirit to Chung and Ely (2003), our argument is quite distinct from theirs.

Our article 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

8. 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 between agents' signals in designing incentives to approximate the Nash equilibrium under complete 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.

9. An undominated Nash equilibrium is a Nash equilibrium in which no player ever uses a weakly dominated action.

10. 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.

subgame-perfect implementation approach of Moore and Repullo (1988).[11] In particular, Maskin and Tirole (1999a), henceforth MT, show that the nonverifiability of states of nature can be overcome by using a three-stage subgame-perfect implementation mechanism that induces truth-telling by all parties as the unique equilibrium outcome, and does so in pure strategies. We contribute to this debate in two ways. First we show that the introduction of even small information perturbations greatly reduces the power of subgame-perfect implementation. This suggests that the introduction of incomplete information can significantly change the insights obtained by MT. Second, we show that when there is asymmetric information ex post about the good's valuation, an outside option for the seller permits a more efficient outcome. We argue that this option can be seen as corresponding to ownership of an asset.

The article is organized as follows. Section II uses a simple buyer-seller example to introduce the MR mechanism, to show why truthful implementation using this mechanism is not robust to small information perturbations, and why such perturbations generate an undesirable sequential equilibrium. Section III extends our analysis to general MR mechanisms with n states of nature and transferable utility, and shows that for a given social choice rule f, truth-telling equilibria are only robust to small information perturbations if this f is *strategy-proof* (which in turn implies Maskin monotonicity under weak assumptions on preferences).[12] In Section IV, we ask whether *any* extensive form mechanism is robust to small information perturbations. There we prove that for any social choice rule that is not Maskin monotonic one can find small information perturbations under which an undesirable sequential equilibrium exists. Section V considers the case of full informational asymmetry ex post and shows that asset ownership, by providing outside options, can lead to approximately efficient ex ante investments, whereas contracts or mechanisms with no outside option cannot. Finally, Section VI concludes with a few remarks and also suggestions for future research.

11. For example, see Aghion, Dewatripont, and Rey (1994) and Maskin and Tirole (1999a, 1999b).

12. If f is strategy-proof, it is always a weakly dominant strategy for each agent to tell the truth in the direct mechanism associated with f. See also Definition 1 for a precise definition of strategy-proofness.

II. Hart-Moore Example of the Moore-Repullo Mechanism

II.A. Basic Setup

Consider the following simple example from Hart and Moore (2003), which captures the logic of Moore and Repullo's (1988) 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 and θ is the good's quality. 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 $\theta_H = 14$, and if it is low quality then B values it at $\theta_L = 10$. 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.

II.B. 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, so no initial contract between the two parties can be made credibly contingent on θ, 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 a "high" or "low" quality. If B announces "high" then B pays S a price equal to 14 and the game 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 $F = 9$ 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 a payment of 6 from B) and the game stops.
    (d) If B rejects at 3b then S pays F to T

(e) B and S each get the item with probability $\frac{1}{2}$.

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, consider first the case $\theta = \theta_H$. 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 price of 6 at stage 3b (since by rejecting he will end up at stage 3e and get $\frac{14}{2} = 7$, but since the good is worth 14 he gets $14 - 6 = 8$ by accepting). Anticipating this, S knows that if she challenges B, she will receive $9 + 6 = 15$, which is greater than 10 that she would receive if she did not challenge. Moving back to stage 1, if B lies and announces "low" when the true state is high, he gets $14 - 9 - 6 = -1$, whereas he gets $14 - 14 = 0$ if he tells the truth, so truth telling is the unique equilibrium here. Truth telling is also the unique equilibrium when $\theta = \theta_L$: in that case S will not challenge B when B (truthfully) announces "low," 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 the lottery which assigns the item randomly instead B can secure a surplus equal to $\frac{10}{2} = 5$). Anticipating this, S will not challenge B because doing so would give her a net surplus equal to $\frac{10}{2} - 9 = -4$ which is less than the payment of 10 she receives if she does not challenge B's announcement.

This mechanism (and more generally, the Moore-Repullo mechanisms we describe in Section III) has two nice and important properties. First, it yields unique implementation in subgame-perfect equilibrium, that is, 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 because it involves telling the truth. In what follows, we show that both of these properties fail once we introduce small information perturbations.

II.C. The Failure of Truth Telling with Perturbed Beliefs about Value

1. Pure Strategy Equilibria. As in the example above, we continue to suppose that the good has possible values $\theta \in \{\theta_H, \theta_L\}$ with

$\theta_H = 14$ (the high state) and $\theta_L = 10$ (the low state). However, we now suppose that the players have imperfect information about θ. Specifically, we suppose they have a common prior μ, with $\mu(\theta_H) = 1 - \alpha$, $\mu(\theta_L) = \alpha$ for some $\alpha \in (0, 1)$, and that each player receives a draw from a signal structure with two possible signals s^h or s^ℓ, where s^h is a high signal that is associated with θ_H, and s^ℓ is a low signal associated with θ_L. We use the notation $s_B = s_B^h$ (resp. $s_B = s_B^\ell$) to refer to the event in which B receives the high signal s^h (resp. the low signal s^ℓ) and similarly we use the notation $s_S = s_S^h$ (resp. $s_S = s_S^\ell$) to refer to the event in which S receives the high signal s^h (resp. the low signal s^ℓ). 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:

(*)

ν^ε	s_B^h, s_S^h	s_B^h, s_S^ℓ	s_B^ℓ, s_S^h	s_B^ℓ, s_S^ℓ
θ_H	$(1-\alpha)\left(1-\varepsilon-\varepsilon^2\right)$	$(1-\alpha)\varepsilon$	$\frac{(1-\alpha)\varepsilon^2}{2}$	$\frac{(1-\alpha)\varepsilon^2}{2}$
θ_L	$\frac{\alpha\varepsilon^2}{2}$	$\frac{\alpha\varepsilon^2}{2}$	$\alpha\varepsilon$	$\alpha\left(1-\varepsilon-\varepsilon^2\right)$

Note that for all ε, the marginal probability distribution of ν^ε on θ coincides with μ, and that as ε converges to 0, ν^ε assigns probability converging to 1 to the signals being correct. Note also that the buyer's signal becomes infinitely more accurate than the seller's signal as $\varepsilon \to 0$. This special feature implies that when deciding whether to challenge the buyer if S and B were informed of both signals, and the signals disagree, they will conclude that with high probability the state corresponds to B's signal.

We now show that there is no equilibrium in pure strategies in which the buyer always reports truthfully. To simplify the exposition of this example, we keep the payments under the perturbed mechanism the same as in the MR mechanism under common knowledge of the previous subsection and assume that B must participate in the mechanism. This is equivalent to assuming that B's participation constraint is slack, which in turn can be arranged by a constant ex ante payment and so does not influence the incentives for truth telling. By way of contradiction, suppose there is a pure strategy equilibrium in which B reports truthfully, and consider B's play when $s_B = s_B^h$. Then B believes that, regardless of what signal player S gets, the expected value of the good is greater than 10. So B would like to announce "low" if he expects that S will not challenge the

announcement. If B does announces "low," then in a fully revealing equilibrium, S will infer that B must have received the low signal, that is, $s_B = s_B^\ell$. But under signal structure $(*)$, S thinks that B's signal is much more likely to be correct, so S now believes that there is a large probability that $\theta = \theta_L$; 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 the high signal s_B^h. Therefore, there does not exist a fully revealing 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. Allowing for Mixed Strategies. The result that there are no truthful equilibria in pure strategies leaves open the possibility that there are mixed strategy equilibria in which the probability of truthful announcement goes to 1 as ε goes to 0. This is close to 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 that this is not the case under the signal structure $(*)$.

Let σ_B^h denote the probability that B announces "low" after receiving the high signal s_B^h, and let σ_B^ℓ be the probability B announces "high" after receiving the low signal s_B^ℓ, as in the following table:

	High	Low
s_B^h	$1 - \sigma_B^h$	σ_B^h
s_B^ℓ	σ_B^ℓ	$1 - \sigma_B^\ell$

The corresponding mixing probabilities for player S are

	Challenge	Don't Challenge
s_S^h	$1 - \sigma_S^h$	σ_S^h
s_S^ℓ	σ_S^ℓ	$1 - \sigma_S^\ell$

Then for mixed strategy equilibria of the mechanism to converge to the equilibrium under complete information where the buyer announces the valuation truthfully, we should have $\sigma_B^{\varepsilon,h}$, $\sigma_B^{\varepsilon,\ell}$, $\sigma_S^{\varepsilon,h}$, and $\sigma_S^{\varepsilon,\ell}$ all converge to 0 as $\varepsilon \to 0$. However, this is not the case, as shown by the following

PROPOSITION 1. Under the information perturbations corresponding to (∗), there is no sequence of equilibrium strategies $\sigma_B^\varepsilon, \sigma_S^\varepsilon$ such that $\sigma_B^{\varepsilon,h}, \sigma_B^{\varepsilon,\ell}, \sigma_S^{\varepsilon,h}$, and $\sigma_S^{\varepsilon,\ell}$ all converge to 0 as $\varepsilon \to 0$.

Proof of Proposition 1. Suppose to the contrary that there is a sequence of equilibrium strategies $\sigma_B^\varepsilon, \sigma_S^\varepsilon$ such that $\sigma_B^{\varepsilon,h}, \sigma_B^{\varepsilon,\ell}, \sigma_S^{\varepsilon,h}$, and $\sigma_S^{\varepsilon,\ell}$ all converge to 0 as $\varepsilon \to 0$. In stage 1, the expected payoff of player B who received the low signal s_B^ℓ and plays "High (H)" tends to -4 while the expected payoff of player B who received the low signal s_B^ℓ and plays "Low (L)" tends to 0 (here, player B makes use of the signal distribution (∗) together with the expectation that the seller's strategies $\sigma_S^{\varepsilon,h}$ and $\sigma_S^{\varepsilon,\ell}$ converge to 0 as $\varepsilon \to 0$, B believes with high probability that S does not "Challenge"). Now, in stage 1, the expected payoff of player B who received the high signal s_B^h and plays "High (H)" tends to 0 while the expected payoff of player B who received the high signal s_B^h and plays "Low (L)" in the limit is below $\max\{14-6-9, 7-9\} = -1$ (recall that B believes with high probability that S chooses "Challenge"). So for ε small, there is no σ that makes player B indifferent between H and L, so player B plays in pure strategies in Stage 1. As in the argument about pure-strategy equilibrium, the fact that B's signal is much more accurate than S's implies that such a strategy profile is not an equilibrium. ∎

This shows that one appealing property of the unique equilibrium in the MR mechanism under common knowledge (namely, a good equilibrium is a truthful one) can disappear once we introduce small information perturbations. In the next subsection we show the nonrobustness of another appealing property of the MR mechanism under common knowledge: that it uniquely implements any desired social choice function.

II.D. 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 information perturbations. But in fact one can go further and exhibit arbitrarily small information perturbations for which the 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 that remains bounded away from the

sequential (or subgame-perfect) equilibrium outcome under complete information.

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

(**)

ν^ε	s_B^h, s_S^h	s_B^h, s_S^ℓ	s_B^ℓ, s_S^h	s_B^ℓ, s_S^ℓ
θ_H	$(1-\alpha)(1-\varepsilon^2)$	$\frac{(1-\alpha)\varepsilon^2}{3}$	$\frac{(1-\alpha)\varepsilon^2}{3}$	$\frac{(1-\alpha)\varepsilon^2}{3}$
θ_L	$\alpha\varepsilon^2$	$\frac{\alpha\varepsilon}{2}$	$\frac{\alpha\varepsilon}{2}$	$\alpha(1-\varepsilon-\varepsilon^2)$

With this signal structure, both agents believe with high probability that if they receive different signals, the signal corresponding to the low state is correct.

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

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, that is, 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 for sequential equilibrium, we also have to assign beliefs over states and signals for each signal of each player and any history of play. Before playing the game but after receiving their private signals, 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 price of 6 (which requires that S will have challenged), we assume that B always believes with probability 1 that the state is θ_L and that S has received the low signal s_S^ℓ.

So what we want to show is that for $\varepsilon > 0$ sufficiently small, the strategy profile is *sequentially rational* given the beliefs we just described and that, conversely, these beliefs are *consistent* given the strategy profile. Here we check sequential rationality (the basic intuition for the belief consistency part of the proof is given in note 13). To establish sequential rationality, we solve the

game backward. At stage 3, regardless of his signal, B believes with probability 1 that the state is θ_L. Accepting S's offer at price of 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 1, her offer at price of 6 will be rejected by B in the next stage, thus S anticipates that as ε becomes small, the payoff is approximately equal to $7-9=-2$ if her signal is high (equal to s_S^h) and to $5-9=-4$ if the signal is low (equal to s_S^ℓ). On the contrary, if S chooses "Not 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 the high signal s_B^h. Then, as ε becomes small, B believes with high probability that the true state is θ_H 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 the low signal s_B^ℓ, it is optimal for her to announce "Low."[13]

As we will see in the next section, the fact that the MR mechanism cannot induce even approximate truth telling under information 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 nonrobustness of subgame-perfect implementation using MR mechanisms, we review Maskin's necessity result on Nash implementation, and explain why the social choice function we try to implement in this example is not Maskin monotonic.

13. 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's 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, the induced sequence of beliefs about θ will converge toward ν^e conditioned on private signals along the equilibrium path of the pure-strategy equilibrium. When B is offered the good at price of 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 trembled 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^ℓ rather than when receiving s_S^h, then B will assign probability close to 1 to S receiving signal s_S^ℓ.

II.E. This Example Does Not Satisfy Maskin Monotonicity

1. Maskin's Necessity Result on Nash Implementation. Recall that a social choice function f on state space Θ is *Maskin monotonic* if for all pair of states of nature (preference profiles) θ' and θ'' if $a = f(\theta')$ and

$$\left\{(i, b)\middle| u_i(a; \theta') \geq u_i(b; \theta')\right\} \subseteq \left\{(i, b)\middle| u_i(a; \theta'') \geq u_i(b; \theta'')\right\}$$

(i.e., no individual ranks a lower when moving from θ' to θ''), then $a = f(\theta'')$. Here $u_i(a; \theta)$ denotes player i's utility from outcome a in state θ. A *social choice function* (SCF) f is said to be *Nash implementable* if there exists a mechanism $\Gamma = (M, g)$ where $m = (m_1, \ldots, m_n) \in M = M_1 \times \ldots \times M_n$ denotes a strategy profile and $g : M \to A$ is the outcome function (which maps strategies into outcomes), *and if* for any θ the Nash equilibrium outcome of that mechanism in state θ is precisely $f(\theta)$. Then, Maskin (1999) shows that if f is Nash implementable, it must be Maskin monotonic.

Let us summarize the proof, which we refer to again later. By way of contradiction, if f were not Maskin monotonic, then there would exist θ' and θ'' such that for any player i and any alternative b

$$u_i(f(\theta'); \theta') \geq u_i(b; \theta') \Longrightarrow u_i(f(\theta'); \theta'') \geq u_i(b; \theta'') \tag{1}$$

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 $m^*_{\theta'}$ of the game $\Gamma(\theta')$. By definition of Nash equilibrium, we must have

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

But then, from (1) we must also have

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

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. The Social Choice Function in Our Example Is Not Maskin Monotonic. It is easy to show that the social choice function in our

Hart-Moore example is not Maskin monotonic. The set of social outcomes (or alternatives) A is defined as:[14]

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

where q is the probability that the good is traded from S to B; y_B, y_S are the transfers of B and S, respectively; and the utility functions of the seller and the buyer are, respectively:

$$u_S(q, y_B, y_S; \theta) = y_S$$

and

$$u_B(q, y_B, y_S; \theta) = \theta q + y_B.$$

The two states of the world are θ_H and θ_L, which correspond respectively to the good being of high and low quality. We have just seen that if an SCF f under which trade occurs with probability 1 is Maskin monotonic, then we must have:

$$f(\theta_H) = f(\theta_L).$$

The SCF we seek to implement requires that

$$f(\theta_L) = (1, -10, 10),$$
$$f(\theta_H) = (1, -14, 14).$$

Clearly $f(\theta_L) \neq f(\theta_H)$, but the buyer ranks outcome (1, −10, 10) at least as high under θ_L as under θ_H, while the seller has the same preferences in the two states. Thus, f is not Maskin monotonic, so Maskin's result implies that this f is not Nash implementable. It is implementable by a MR mechanism under common knowledge, but it is not implementable by this mechanism under information perturbations.

Our analysis in the next two sections is motivated by the following questions. (1) Is the nonexistence of truth-telling equilibria in arbitrarily small information 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 equilibria in arbitrarily small information perturbations of the above MR mechanism, directly linked to f being non–Maskin monotonic?

In Section III, we consider a more general version of the MR mechanism and link the failure of MR mechanisms to implement

14. The sum $y_S + y_B$ can be negative to allow for penalties paid to a third party.

truth telling in equilibrium under information perturbations to the lack of Maskin monotonicity of the corresponding SCF. Then in Section IV, we consider any sequential mechanism that implements a non–Maskin monotonic SCF (and more generally, *social choice correspondences*, SCC) under common knowledge, and show that for an arbitrarily small information perturbation of the game 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.

III. 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, Moore and Repullo (1988) consider environments where utilities are transferable and show that truth telling is a unique subgame-perfect equilibrium in the MR mechanisms. Since this is the most hospitable environment for subgame-perfect implementation, and because most contracting settings are in economies with money, we focus on it.

III.A. 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 $\Theta_i = \{\theta_i^1, \ldots, \theta_i^n\}$ for each $i = 1, 2$.[15] 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.[16] Preference characteristics (θ_1, θ_2) are common knowledge between the two parties but not verifiable by a third party.

15. Moore and Repullo (1988) allow for an infinite state space but impose bounds on the utility functions.

16. Because we do not assume that the prior on Θ is a product measure, the product structure of $\Theta = \Theta_1 \times \Theta_2$ is not crucial to our results. To see this, note that given any finite set of states of nature Θ and utility functions $u_i : \Theta \times A \to \mathbb{R}$ for each

Let $f=(D, T_1, T_2)$ be an SCF 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 $(t_1, t_2)=(T_1(\theta_1, \theta_2), T_2(\theta_1, \theta_2))$.

Moore and Repullo (1988) propose the following class of mechanisms. These mechanisms involve two phases, where phase i is designed 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 θ_1 and then carries on with phase 2 in which player 2 announces θ_2. Phase 1 proceeds as follows:

(1) Player 1 announces a preference θ_1, and we proceed to stage 2.
(2) If player 2 announces ϕ_1 and $\phi_1=\theta_1$, then phase 1 ends and we proceed to phase 2. If player 2's announcement ϕ_1 does not agree (i.e., $\phi_1 \neq \theta_1$) then player 2 "challenges" and we proceed to stage 3.
(3) Player 1 chooses between

$$\{x; t_x + \Delta\}$$

and

$$\{y; t_y + \Delta\},$$

where $x=x(\theta_1, \phi_1)$ and $y=y(\theta_1, \phi_1)$ depend on both θ_1 and ϕ_1 and Δ is a positive number suitably chosen (see below) and (x, y, t_x, t_y) are 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, this corresponds to the buyer refusing the offer at price 6), then player 1 pays $t_1=t_x+\Delta$ and 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 Hart-Moore example, this corresponds to the buyer taking up the offer at price 6),

player i, we can identify Θ_i with the collection of $\{u_i(\cdot, \theta)\}_{\theta \in \Theta}$. Now, define $\tilde{u_i}: \Theta_1 \times \Theta_2 \times A \to \mathbb{R}$ as follows: for $\theta_i=u_i(\cdot, \theta)$ we set $\tilde{u_i}(\cdot, \theta_i) := u_i(\cdot, \theta)$. This setting is equivalent to the former one.

then player 1 pays $t_1 = t_y + \Delta$ and player 2 receives $t_2 = t_y + \Delta$. The game ends here.

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 argument applies as follows when the state of nature θ is common knowledge: If player 1 lies at stage 1.1, then player 2 will challenge, and 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 1 will find it optimal to announce the truth and thereby implement the SCF 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.

III.B. Perturbing the Information Structure

We now show that this result does not hold for small perturbations of the information structure of the following form: each agent $i = 1, 2$ receives a signal $s_i^{k,l}$ where k and l are both integers in $\{1, \ldots, n\}$; 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 and signal pairs (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). In these expressions, we abuse notation and write: $\mu(\theta_1^k, \theta_2^\ell)$ for the $\text{marg}_\Theta(\mu)(\theta_1^k, \theta_2^\ell)$. This corresponds to an information perturbation such that each player i's signal is much more informative about his own preferences

than about those of the other player. Note that in an intuitive sense the prior ν^{ε} is close to $\mu$when ε is small; this is also true in a formal sense.[17]

We begin by considering pure strategy equilibria. For this purpose, we make use of the concept of strategy-proofness:

DEFINITION 1. An 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, an SCF f is strategy-proof if telling the truth is a weakly dominant strategy through a direct mechanism associated with f whereby the players are asked to announce their preference parameter. Strategy-proofness implies a weak version of Maskin monotonicity, namely, that for any θ, θ' such that

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

we have $f(\theta) = f(\theta')$.[18] As a corollary, strategy-proofness also implies the usual Maskin monotonicity condition when preferences over outcomes in $f(\Theta)$ are strict, where $f(\Theta)$ denotes the range of f.

17. For concreteness we specify the supremum-norm topology when discussing the convergence of the priors. That is, let $\mathcal{P}$ denote the set of priors over $\Theta \times S$ with the following metric $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$: for any μ, $\mu' \in \mathcal{P}$,

$$d(\mu, \mu') = \max_{(\theta,s)\in\Theta\times S} |\mu(\theta, s) - \mu'(\theta, s)|.$$

So, when we say $\nu^k \to \mu$, we mean that $d(\nu^k, \mu) \to 0$ as $k \to \infty$.

18. 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 condition. 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 therefore the argument yields the usual Maskin monotonicity condition. Our weak monotonicity is closely related to conditions proposed by Dasgupta, Hammond, Maskin (1979). In that paper, strategy-proof SCFs are characterized via the concept of "independent person-by-person monotonicity" which is stronger than our condition of weak Maskin monotonicity.

THEOREM 1. Suppose that a non-strategy-proof SCF f is implementable by an MR mechanism under complete information. 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.

Proof of Theorem 1. Under the signal structure $(* * *)$, if player 2 sees that player 1's announcement about θ_1 is different from her signal, and she believes player 1 is reporting "truthfully," she disregards her own information on Θ_1 and follows player 1's announcement (and symmetrically for player 1 vis-à-vis player 2 regarding signals over Θ_2).

Now, suppose that f is not strategy-proof. Then there is a player, say player 1, and states $\theta_1^h, \theta_1^k, \theta_2^l$ such that

$$u_1(f(\theta_1^h, \theta_2^l); \theta_1^h) < u_1(f(\theta_1^k, \theta_2^l); \theta_1^h).$$

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 $s_1^{h,l}$ and player 2 gets signal $s_2^{h,l}$. Player 1 would like to announce "θ_1^k" if she expects that subsequent to such an announcement, player 2 agrees with "θ_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 (θ_1^k, θ_2^l). Consequently, player 2 will not challenge player 1's announcement. But then, anticipating this, player 1 will announce "θ_1^k" and thereby receive $f(\theta_1^k, \theta_2^l)$ instead of $f(\theta_1^h, \theta_2^l)$. This in turn shows that there does not exist a truthfully revealing equilibrium in pure strategies. ∎

Theorem 1 links the nonrobustness of the MR mechanism to the failure of Maskin monotonicity of the SCF to be implemented. For instance, in the Hart-Moore example in Section II, the SCF is not Maskin monotonic and preferences over $f(\Theta)$ are strict, so the SCF in that example is not strategy-proof.

Note that the foregoing 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 information perturbations. Moreover, the 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.

Next, we turn attention to mixed-strategy equilibrium. If we require that both players tell (at least, approximately) the truth in each of the two phases, then *no* SCF $f=(D, T_1, T_2)$ 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 information perturbations.

More formally, in the Online Appendix we prove the following.

THEOREM 2. Suppose that an SCF f is implementable by an MR mechanism under complete information. 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.

Here is an intuition for why requiring approximate truth telling by both players in each phase precludes robust implementation by the MR mechanism. Suppose that both players receive a signal that 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 $\hat{\theta}$, then player 2 (playing 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 $\hat{\theta}$ and so will not follow her private signal and thus she is not reporting truthfully.

Let us make two remarks at this stage. First, the nonrobustness of truth telling as a sequential equilibrium outcome of the MR mechanism is of interest because truth telling is cognitively simple, and also because the nonexistence of a truthful sequential equilibrium implies the nonexistence of a desirable pure equilibrium, and implementation theory has mainly focused on pure-strategy equilibria. Second, neither of the nonrobustness results

of this section rule out the possibility that some SCF f can be implemented as the limit of mixed-strategy (nontruthful) sequential equilibrium outcomes.[19] However, in the next section, we show that if f is not Maskin monotonic but can be implemented by the MR or by any other extensive form mechanism under common knowledge, then there always exist arbitrarily small information perturbations under which there also exist sequential equilibria with undesirable outcomes.

IV. Any Mechanism

In this section, we go beyond MR mechanisms and consider the set of all extensive form mechanisms. Suppose a non–Maskin monotonic SCF is implemented by a (not necessarily MR) mechanism under complete information. Then, we show that there always exists a "bad" sequential equilibrium in arbitrarily small information perturbations of that mechanism. We begin by presenting the argument in a nutshell, using the Hart-Moore example to illustrate our point. Finally, we proceed to state and establish a more general result that covers SCCs as well as SCFs.

IV.A. 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 introducing just a small amount of incomplete information markedly enlarges the set of (sequential) beliefs that are consistent with Bayesian rationality. As a result, one can turn an arbitrary Nash equilibrium of an extensive form mechanism that implements a non–Maskin monotonic SCF f under common knowledge into a sequential equilibrium of the perturbed game.

More specifically, suppose there are n players, and each player i has a state dependent 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 $\Gamma(\theta)$; let $SPE(\Gamma(\theta))$ denote the set of subgame-perfect equilibria of the game $\Gamma(\theta)$. An SCF f is said to be subgame-perfect

19. For conditions under which the unique subgame-perfect equilibrium outcome of a perfect information game remains an equilibrium outcome in perturbed games, see Takahashi and Tercieux (2011).

implementable if there exists a mechanism $\Gamma = (M, g)$ such that for each state θ, every subgame-perfect equilibrium outcome coincides with $f(\theta)$. Here is an informal statement of the main result.

1. Main Result. Assume finite state space and finite strategy spaces.[20] Assume, further, that a mechanism Γ subgame-perfect implements a non–Maskin monotonic SCF f under complete information. Then there exists a sequence of information 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 do not converge to $f(\theta)$ in some state θ as $\varepsilon \to 0$.

In particular, under the usual additional conditions where Maskin monotonicity is sufficient for Nash implementation, this result implies the following: whenever an 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 information perturbations.

2. Intuition for the Proof. Suppose that the SCF f is not Maskin monotonic. Then, there exist θ' and θ'' such that for any player $i \in N$ and any alternative $b \in A$

$$u_i(f(\theta'); \theta') \geq u_i(b; \theta') \Longrightarrow u_i(f(\theta'); \theta'') \geq u_i(b; \theta'') \tag{2}$$

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 state θ' such that $g(m_{\theta'}) = f(\theta')$. But then using the same argument as in the proof of Maskin's theorem summarized in Section II, $m_{\theta'}$ 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 information 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 II showed the nonrobustness of the particular MR mechanism considered there.

20. In the Online Appendix we extend the result to the case of countable strategy sets.

As a concrete example, consider again the MR mechanism studied in Section II. Under common knowledge of the state, it is a Nash equilibrium for B to announce θ_L at stage 1 and for S to never challenge at stage 2. However, this is a bad Nash equilibrium and it is "not" a sequential equilibrium. 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 received the low signal s_S^ℓ, 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 the low state θ_L if B "knew" that the state was θ_H, but this can become consistent under the perturbation. This is the key to how the perturbation turns a bad (non-sequential) Nash equilibrium of the game with complete information into a sequential equilibrium in the perturbed game.

IV.B. 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 Online Appendix for the formal proof. In the first reading, the reader can skip the rest of Section IV here and go directly to Section V without losing much of the main idea.

1. The Environment. In what follows, we consider a more general environment, with a finite set $N = \{1, \ldots, n\}$ 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 quasi-linear preferences with transferable money, as was needed for MR mechanisms. Each player i has a state-dependent utility function $u_i : A \times \Theta \to \mathbb{R}$, where Θ is a finite set of states of nature.[21]

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

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 $s = (s_1, \ldots, s_n) \in S \equiv \times_{i \in N} S_i$. 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 write $\mu(\cdot|s_i)$ 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 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 assume for each i and θ, the marginal distribution on i's signals places strictly positive weight on each of i's signals in every state, that is, $\mu(s_i^\theta) \equiv [\text{marg}_{S_i}\mu](s_i^\theta) > 0$, so that Bayes's rule is well defined. Note that in case μ is a complete information prior, this implies in particular that for each $(\theta, s^\theta) \in \Theta \times S : \mu(\theta, s^\theta) > 0$.

An SCC is a set-valued mapping $\mathcal{F} : \Theta \rightrightarrows A$. We have focused on SCFs in the previous sections. In this section, we generalize our arguments to encompass SCCs.

Since we consider more general extensive form mechanisms than MR mechanisms, we need to introduce some notation. Most of the notation used here is consistent with Moore and Repullo (1988). The reader is referred to that paper for the definition and notation of extensive form mechanisms. We restrict attention to mechanisms that are multistage games with observed actions, meaning 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.[22] We also assume that the mechanism has a finite number of stages. The class of mechanisms we consider in the present paper is exactly the same as the one Moore and Repullo (1988) allowed. A *mechanism* is then an extensive game form $\Gamma = (\mathcal{H}, M, \mathcal{Z}, g)$ where (1) $\mathcal{H}$ is the set of all histories; (2) $M = M_1 \times \ldots \times M_n$, $M_i = \times_{h \in \mathcal{H}} M_i(h)$ 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

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

function that maps the set of terminal histories (denoted H_T) into the set of outcomes (A).

The following notation will be useful: An element of $M(h) = M_1(h) \times \ldots \times M_n(h)$, say $m(h) = (m_1(h), \ldots, m_n(h))$ 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.

There is an initial history $\emptyset \in \mathcal{H}$, and $h_t =: (\emptyset, m^1, m^2, \ldots, m^{t-1})$ is the history at the end of period t, where for each k, $m^k \in M(h_k)$. If for $t' \geq t+1$, $h_{t'} = (h_t, m^t, \ldots, m^{t'-1})$, then $h_{t'}$ follows history h_t. As Γ contains finitely many stages, there is a set of terminal histories[23] $H_T \subset \mathcal{H}$ 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 also denote $g(m; h)$ the outcome that obtains when players use strategy profile m starting from history h, that is, $g(m; h) = g(h_T[m, h])$. In what follows, we only consider finite mechanisms.

ASSUMPTION 1. $M_i(h)$ is finite for each i and h.

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 Online Appendix, we provide additional assumptions on the class of mechanisms so that our result can be extended to countable message spaces. This extension is important because the literature often uses integer games (i.e., games where one

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

dimension of the message space is the set of positive integers) as part of implementing mechanisms.[24]

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 $h \in \mathcal{H} \backslash H_T$. Recall that $SPE(\Gamma(\theta))$ denotes the set of subgame-perfect equilibria of the game $\Gamma(\theta)$ and $NE(\Gamma(\theta))$ denotes 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. In what follows, whenever players face uncertainty about the state and other player's signals, they possess a probabilistic belief over this uncertainty and with respect to this belief, they aim to maximize expected utility.[25] A strategy profile $\sigma = (\sigma_1, \ldots, \sigma_n)$ lists a strategy for each player i where $\sigma_i : S_i \to M_i$ and $\sigma_i(h_t, s_i)$ is a 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, that is., a function that maps the set of possible histories and signals into the set of probability distributions over messages: $\sigma_i(\cdot|h_t, s_i) \in \Delta(M_i(h_t))$ is the probability distribution over $M_i(h_t)$ given history h_t and signal s_i.

With this notation in place we can restate the definition of *sequential equilibrium* as specialized to these multistage games of observed actions. A sequential equilibrium is a profile of assessment (or beliefs) ϕ and strategies σ that satisfy both

24. Our results do not critically depend on the countability assumption. We believe that our results would hold for arbitrary mechanisms if we were to use perfect Bayesian equilibrium (Fudenberg and Tirole 1991b) instead of sequential equilibrium as the solution concept.

25. All the results extend to more general representations for preferences under uncertainty. The interested reader is referred to Kunimoto and Tercieux (2009) for details.

consistency and *sequential rationality*. Here consistency is the requirement that there exists a sequence of totally mixed strategy profiles σ^n converging to σ such that the beliefs ϕ^n computed from σ^n using Bayes's rule converge to ϕ. Sequential rationality means that for each period t and history h^{t-1} up to $t-1$, the continuation strategies are optimal for each player i given the opponents' strategies and his belief ϕ_i. A more formal definition of sequential equilibrium can be found in the Online Appendix.

2. The Existence of a Bad Sequential Equilibrium with Almost-Perfect Information. Although we already introduced the definition of Maskin monotonicity for social choice functions in Section II, we need to extend it to social choice correspondences. A social choice correspondence $\mathcal{F}$ on a payoff relevant state space Θ is *Maskin monotonic* if for all pair of states of nature θ' and θ'' if $a \in \mathcal{F}(\theta')$ and

$$(3) \qquad \{(i,b)\big|u_i(a;\theta') \geq u_i(b;\theta')\} \subseteq \{(i,b)\big|u_i(a;\theta'') \geq u_i(b;\theta'')\}$$

(i.e., no individual ranks a lower when moving from θ' to θ'') then $a \in \mathcal{F}(\theta'')$. We are now in a position to provide a more formal statement of our main theorem.

THEOREM 3. ASSUME ASSUMPTION 1. Suppose that a mechanism SPE implements a non–Maskin monotonic SCC $\mathcal{F}$ and suppose that A is a Hausdorff space[26]. 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 equilibrium assessments and strategy profiles $\{(\phi^\varepsilon, \sigma^\varepsilon)\}_{\varepsilon>0}$ such that as ε tends to 0, $g(\sigma^\varepsilon(s^\theta); \emptyset) \to a \notin \mathcal{F}(\theta)$ for some $\theta \in \Theta$ and some outcome $a \in A$.

Proof. See Online Appendix. ∎

REMARK 2. The essence of the proof is to show by construction that if a mechanism implements by subgame-perfect equilibrium alternative a for state θ', and if $\{(i, b)|u_i(a;\theta') \geq u_i(b;\theta')\} \subseteqq \{(i, b)u_i(a;\theta'') \geq u_i(b;\theta'')\}$, then there is a sequence of priors converging to the complete-information prior and a corresponding sequence of sequential equilibria of this mechanism such

26. That is, a topological space in which any two distinct points can be separated by two disjoint open sets. For example, R^n with the usual topology is a Hausdorff space.

that the conditional probability of a given θ'' goes to 1. This shows that whenever an SCC cannot be implementable using a static mechanism due to the violation of Maskin monotonicity, this SCC cannot be implemented using an extensive form mechanism that is robust to the introduction of a small amount of incomplete information.

REMARK 3. While non–Maskin monotonic SCFs cannot be robustly implemented, things are quite different for Maskin monotonic SCFs. Here we restrict our focus to SCF's rather than SCCs. In the Online Appendix we extend the argument to the case of SCCs.

What appears as a natural candidate for "robust implementation" of a SCF amounts to constructing a Nash implementable mechanism with the following two properties: (1) there exists at least one strict Nash equilibrium; and (2) the map from information structures to Nash equilibria has a closed graph, so adding a small amount of incomplete information only slightly increases the set of Nash equilibria. In the Online Appendix, we formalize these two properties and propose a definition of robust Nash implementation.

To see this, note that the first property ensures that the strict Nash equilibrium continues to be a strict (Bayesian) Nash equilibrium for any nearby environment and hence that there is always a good equilibrium for any nearby environment. The second property in turn ensures that all Nash equilibria will continue to have outcomes that are close to the desired outcome for any nearby environment.

Regarding the first property, the existence of a strict Nash equilibrium in a mechanism that implements an SCF can easily be ensured under a slight strengthening of Maskin monotonicity, namely, strong Maskin monotonicity. In the Online Appendix, we show that this is also the case for SCCs.

As to the second property, in many situations, Nash implementation of Maskin monotonic SCFs can be achieved using finite mechanisms (see Saijo 1988). Routine arguments then imply that the second property is satisfied.[27]

27. This property comes from the following two facts. First, a small change in the prior probability corresponds to a small change in ex ante payoffs. Second, the pure Nash equilibrium correspondence is upper hemi continuous in the space of payoffs.

For the case of infinite mechanisms, the argument is relegated to the Online Appendix, which provides sufficient conditions under which one can ensure that properties (1) and (2) are satisfied. There we take care of SCCs as well as SCFs. Interestingly, these sufficient conditions are satisfied by any Maskin monotonic SCF in quasi-linear environments with money.

V. Outside Options and the Hold-up Problem

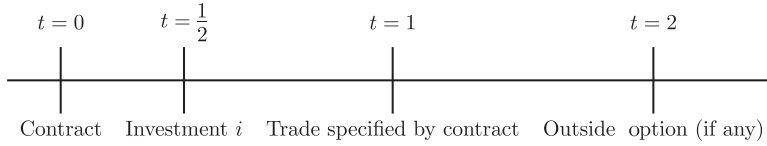
Thus far, we have shown that the mechanisms used by proponents of the "implementation critique" of the property right theory of the firm (e.g., Maskin and Tirole 1999a) are themselves not robust to small deviations from perfect information and common knowledge. That leaves open the question of what role outside options (e.g., as induced by asset ownership as in Grossman and Hart 1986) can play in alleviating the hold-up problem in situations that depart more significantly from complete or just symmetric information.

As a first step in this direction, we consider an environment with an ex ante investment stage and where ex post bargaining takes place under one-sided asymmetric information. We present an example where the presence of an outside option allows mechanisms that approximate ex ante efficiency. Moreover, we argue that static or sequential mechanisms without an outside option cannot do as well, which, in turn, we see as a justification for the role of ownership allocation in contracting under incomplete information.

V.A. The Set-up

Suppose there is a buyer (B) and a seller (S) of a single unit of an indivisible object with utility $\tilde{v}$ to the buyer, where $\tilde{v} \in \{\underline{v}, \bar{v}\}$ and $\bar{v} > \underline{v} > 0$. The utility of the seller for the object is assumed to be always zero. Time is discrete, with a contracting period 0 where the good is offered to the buyer at a prespecified price, an investment period $\frac{1}{2}$ whereby the seller can increase the buyer's valuation for the good; and a trading period 1. Investment is unobservable as in Grossman and Hart (1986). Moreover, we allow for the possibility that an outside option can be exerted in period 2 by one party if trade does not occur in period 1 and focus attention on the case where the outside option yields utility $\underline{v}$ to whoever has the good at that point. A natural interpretation is that $\bar{v}$ is the value the buyer and the seller can generate in their

relationship and $\underline{v}$ is the default value that can be generated outside of the relationship. The timing of the events is as follows:



The seller may make an investment in period $\frac{1}{2}$ that increases the probability that the good is high quality, as in Che and Hausch (1999). Specifically, suppose that at cost $c(i)$ the seller achieves $v = \bar{v}$ with probability i, where $c(\cdot)$ is continuous, twice differentiable, and satisfies $c'(i) > 0$, $c''(i) > 0$, $c(0) = 0$, $c'(1) = +\infty$, and $c'(0) < \bar{v} - \underline{v}$. The buyer will know the value of the good at the beginning of period 1, while the seller will not, so there is one-sided asymmetric information.

V.B. Outside Options as Ownership

One can relate the outside option to the idea of ownership by taking the owner of the good to be the party with the right to exercise the outside option. Thus, under seller ownership, if the seller makes an offer to the buyer but the buyer refuses the offer, then the seller can always choose to always exert his outside option and gets $\underline{v}$.

This interpretation as ownership is consistent with other works in the property rights literature, starting with Grossman and Hart (1986), where ownership of the assets of a firm allows the owner to make alternative use of these assets in case of disagreement in the ex post bargaining with the other party(ies). This in turn enhances the owner's ex post bargaining power, and therefore it increases the fraction of the ex post production surplus the owner can secure in this bargaining, which, in turn, enhances the owner's investment incentives. In our setting too, ownership of the good will allow the seller to extract a higher price from a high-valuation buyer, and anticipating this, the seller will invest a higher i in the relationship. However, as we will show, no mechanism (contract) without an outside option can do as well as a contract with outside option to the seller in inducing efficient investment by the seller in period $\frac{1}{2}$.[28]

28. Work in progress by Bester and Münster (2012) makes a similar point about the value of outside options in a closely related model of performance evaluation.

V.C. *Ex Ante Efficiency and Outside Options*

Under our assumptions, the ex ante efficient outcome is to trade whenever the good is high quality, consume the outside option when the good is low quality,[29] and set investment equal to i^*, where $i^* \in (0, 1)$ is the solution to the following first-order condition:

$$\bar{v} - \underline{v} = c'(i^*).$$

The resulting total surplus is then

$$W^* = i^*\bar{v} + (1 - i^*)\underline{v} - c(i^*).$$

We show how a mechanism with an outside option can come arbitrarily close to this payoff.

In this setting, a mechanism takes as input the buyer's announced value for the good, and specifies a trade probability q, transfers y_S and y_B to the seller and buyer respectively, a probability z_S that the seller gets to keep the good if there is no trade, a probability z_B that the buyer gets the good in that case, and therefore the probability $1 - z_B - z_S \geq 0$ that the good is destroyed when it is not traded (the mechanism does not specify an investment level, nor condition other outcomes on it, as investment is not observable). Thus the mechanism maps the buyer's announcement $\tilde{v} \in \{\underline{v}, \bar{v}\}$ into A where $A = \{(q, y_B, y_S, z_B, z_S) \in [0, 1] \times \mathbb{R}^4_+ |\ y_S + y_B \leq 0,\ z_B + z_S \leq 1\}$. In what follows, we consider the case $z_S \equiv 1$ (so that the seller gets the outside option whenever there is no trade, regardless of the buyer's announcement), and therefore the mechanism boils down to a mapping $f(\tilde{v})$ such that $f(\underline{v}) = (\underline{q}, \underline{y}_B, \underline{y}_S)$ (when the buyer announces $\underline{v}$) and $f(\overline{v}) = (\bar{q}, \bar{y}_B, \bar{y}_S)$ (when the buyer announces $\bar{v}$).

Given that $z_S \equiv 1$, for $\varepsilon > 0$ small enough, the mechanism that implements $(1, -(\bar{v} - \varepsilon), \bar{v} - \varepsilon))$ when the buyer announces valuation $\bar{v}$, and $(0, 0, 0)$ when the buyer announces $\underline{v}$ satisfies incentive compatibility (it is a strictly dominant strategy for the buyer to report her valuation v truthfully), individual rationality, and ex post efficiency, that is, trade occurs if and only if there are social gains from trade.

Now suppose that the buyer and the seller agree on this mechanism with the outside option $\underline{v}$ allocated to the seller at

29. From the viewpoint of social welfare it does not matter which party gets to use the outside option.

the contracting stage. Then, moving back to time $t = \frac{1}{2}$, the seller chooses the level of investment to maximize

$$i(\bar{v} - \varepsilon) + (1 - i)\underline{v} - c(i).$$

Given our assumptions, the optimal investment level i^* (for $\varepsilon > 0$ small enough) is determined by the first-order condition:

$$\bar{v} - \varepsilon - \underline{v} = c'(i^*).$$

From the concavity of the problem, this is approximately the same as the first-best investment when ε is small. Thus, a simple contract with seller's ownership can exactly implement an outcome whose total surplus is arbitrarily close to the first best level; this is what we will mean by "approximate ex ante efficiency."

V.D. Ex Ante Efficiency Cannot Be Approximated without Outside Options

As in the complete information case, a crucial question is: what exactly can be achieved with contracts/mechanisms that do not use outside options, so that $z_S = z_B = 0$? Below, we show that under buyer's private information, any "outside-option-free" contract between the buyer and the seller leads to an outcome that remains bounded away from ex ante efficiency.

First, note that if an SCF f that maps the true buyer's valuation $\tilde{v}$ onto a triplet $f(\tilde{v}) = (\tilde{q}, \tilde{y}_B, \tilde{y}_S)$, and yields zero continuation utility to both parties if trade does not occur, is to be implemented by some (static or sequential[30]) mechanism in Bayesian Nash equilibrium, it must be at least weakly incentive compatible for the buyer to report truthfully. It is simple to show that f is incentive compatible if and only if

$$\underline{v}(\bar{q} - \underline{q}) \leq \underline{y}_B - \bar{y}_B \leq \bar{v}(\bar{q} - \underline{q}). \tag{4}$$

Below we prove that one cannot find SCFs with $z_S = z_B = 0$ that are incentive compatible and approximately ex ante efficient. To show this, suppose to the contrary that for any $\varepsilon > 0$ there is an incentive compatible mechanism f^ε whose ex ante total surplus is at least $W^* - \varepsilon$. Then, the associated probabilities i^ε of high quality

30. Approximate ex ante efficiency cannot be achieved by virtual implementation either, since incentive compatibility is also necessary for virtual implementation to work. But precisely we show that without outside options, one cannot find SCFs that are both approximately ex ante efficient and incentive compatible.

must converge to i^*, the probabilities of trade $\underline{q}^\varepsilon$ and $\bar{q}^\varepsilon$ must both converge to 1, and the difference in transfers (i.e., money "burnt") $|\underline{y}_S^\varepsilon - \underline{y}_B^\varepsilon|$ and $|\bar{y}_S^\varepsilon - \bar{y}_B^\varepsilon|$ must both converge to 0. The incentive compatibility condition (2) then implies that $|\underline{y}_B^\varepsilon - \bar{y}_B^\varepsilon| \to 0$, and this, plus the fact that both $|\underline{y}_S^\varepsilon - \underline{y}_B^\varepsilon|$ and $|\bar{y}_S^\varepsilon - \bar{y}_B^\varepsilon| \to 0$, implies that $|\bar{y}_S^\varepsilon - \underline{y}_S^\varepsilon| \to 0$ as well.

Moving back to time $t = \frac{1}{2}$, the seller will choose investment i to maximize

$$i\bar{y}_S^\varepsilon + (1-i)\underline{y}_S^\varepsilon - c(i) = \underline{y}_S^\varepsilon + i(\bar{y}_S^\varepsilon - \underline{y}_S^\varepsilon) - c(i).$$

Because $|\bar{y}_S^\varepsilon - \underline{y}_S^\varepsilon| \to 0$ and $c' > 0$, the solution i^ε to this problem converges to 0, so investment falls far short of the first-best level, which is not consistent with the assumption that the ex ante total surplus converges to W^*. We conclude that ex ante surplus must be bounded away from efficiency.

This shows that in our example no approximately ex ante efficient SCF can be implemented by a mechanism that does not include an outside option (or some other change to the economic environment).[31] Because approximately efficient outcomes can be implemented when outside options are available, and outside options can be interpreted as resulting from ownership allocation, our results combined provide a justification for the role of ownership allocation in contracting under incomplete information.

V.E. Summary

Analyzing the hold-up problem in a setting with ex post asymmetric information, as we have done in this section, yields an interesting new insight: outside options such as those induced by asset ownership can help relax incentive compatibility constraints and thereby improve ex ante efficiency compared to what can be achieved through "ownership-free" contracts/mechanisms.

31. Schmitz (2002) proves a related impossibility result in an example featuring bilateral trade with only two possible investment levels; Bester and Krähmer (2012) extend this to the case where the seller's action is observable but not verifiable.

VI. Concluding Remarks

We conclude by making a few additional remarks. First, the bad sequential equilibria in Section IV survives a standard equilibrium selection criterion. Cho (1987) defines *forward induction equilibrium*, which is an extension of the Cho and Kreps (1987) *intuitive criterion* in signaling games to more 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 reached only by "bad" deviations, provided that other nodes in h can be reached by nonbad deviations. Here, "bad deviations" are deviations with the following property: suppose that at any information set where 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 Hart-Moore example developed in Section II, we can show that "Challenge" 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_H may occur with positive probability. Thus we can always pick an appropriate belief (for instance, one that would assign probability 1 to θ_H) 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 "Challenge" by the seller makes her strictly better off compared to the equilibrium, proving that "Challenge" cannot be a bad deviation.

Our second remark is that the nonrobustness of subgame-perfect implementation does not mean that implementation is hopeless, but suggests that we should further explore the implications of Nash implementation. It is well known that in many important contexts, Nash implementation (or Maskin monotonicity) is quite demanding. For instance, a well-known result by Muller and Sattherthwaite (1977) states that any onto and ex post efficient SCF defined on the domain of all strict preferences is dictatorial when there are at least three outcomes. Maskin (1999) shows that with only two players, this result extends to SCCs. However, it has also been shown that under some mild domain restrictions, for any SCF f, there is a stochastic social choice function that puts probability close to one on the same outcomes as f and that is Maskin monotonic (see Abreu and Sen

1991 and Matsushima 1988 for the details of this approach).[32] Indeed, we saw that the SCF f we sought to implement in this Hart-Moore example was not Maskin monotonic since $f(\theta_L) = (1, -10, 10) \neq f(\theta_H) = (1, -14, 14)$, and therefore not Nash implementable. However, the ε-approximation of that SCF defined by

$$f^{\varepsilon}(\theta_L) = (1 - \varepsilon, -10, 10) \neq f(\theta_H) = (1 - \varepsilon, -14, 14),$$

is Maskin monotonic since for example, B strictly prefers $(1 - \varepsilon, -10, 10)$ to $(1, -10 - 11\varepsilon, 10)$ when $\theta = \theta_L = 10$ but the reverse is true when $\theta = \theta_H = 14$. Hence, even if f is not Maskin monotonic and therefore not Nash implementable, we can find an ε-close stochastic SCF that is Maskin monotonic and therefore Nash implementable for instance in the Moore and Repullo setting.[33] However, the stochastic nature of this mechanism is problematic in terms of *renegotiation-proofness*. For example, if we consider the SCF f^{ε}: with probability ε, the planner must induce a bad outcome under which trade does not occur.[34] Given that there are gains from trade, agents will definitely have incentives to renegotiate. If this possibility is explicitly taken into account by the contracting parties, then the SCF is not going to be Nash implementable anymore. Thus, stochasticity (or randomness) can help robustly implement nearby efficient SCFs but also raises serious renegotiation-proofness issues.

Finally, we feel that laboratory experiments can be useful in assessing the importance of the effect of information perturbations on the likelihood that truth telling will still occur in

32. Here preferences are defined on lotteries over outcomes and agents are assumed to be expected utility maximizers, so typically the restrictions to domains of strict preferences in Muller and Satterthwaite (1977) or in Maskin (1999) are not going to be satisfied.

33. Note that in the Moore-Repullo setting (i.e., with quasi-linear utilities and arbitrary large transfers), for any SCF f, we have the existence of a bad outcome (i.e., an outcome which, in 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. The stochastic approximation of f can therefore be implemented with a canonical Maskin mechanism, although since the mechanism uses integer games it is less appealing than the simple MR mechanism.

34. Renegotiation is less problematic in the case of "exact" Nash implementation since renegotiation then only occurs *out of equilibrium*.

equilibrium. Preliminary work by Aghion et al. (2009) suggests that the effect is potentially large.[35]

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at QJE online (qje.oxfordjournals.org).

HARVARD UNIVERSITY AND CIFAR
HARVARD UNIVERSITY, UNIVERSITY OF NEW SOUTH WALES
HITOTSUBASHI UNIVERSITY
PARIS SCHOOL OF ECONOMICS

REFERENCES

Abreu, D., and H. Matsushima, "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60 (1992), 993–1008.
Abreu, D., and A. Sen, "Virtual Implementation in Nash Equilibrium," *Econometrica*, 59 (1991), 997–1021.
Aghion, P., M. Dewatripont, and P. Rey, "Renegotiation Design with Unverifiable Information," *Econometrica*, 62 (1994), 257–282.
Aghion, P., E. Fehr, R. Holden, and T. Wilkening *Subgame Perfect Implementation: A Laboratory Experiment*. (Mimeo, 2009).
Aghion, P., D. Fudenberg, and R. Holden. *Subgame Perfect Implementation With Almost Perfect Information*. (NBER Working Paper 15167, 2009).
Aliprantis, C., and K. Border. *Infinite Dimensional Analysis,* 2nd ed. (Berlin: Springer Verlag, 1999).
Bester, H., and D. Krähmer, "Exit Options in Incomplete Contracts with Asymmetric Information," *Journal of Economic Theory*, 147 (2012), 1947–1968.
Bester, H., and J. Münster *Subjective Evaluation versus Public Evaluation*. (unpublished working paper, 2012).

35. Aghion et al. (2009) conduct a laboratory experiment testing the robustness of a Moore-Repullo mechanism to information perturbations. The experiment is meant to mimic the Hart-Moore example spelled out in Section II. Subjects are randomly allocated to the buyer and seller roles, and play the mechanism 10 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 10 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.

Cabrales, A., and R. Serrano, "Implementation in Adaptive Better-Response Dynamics: Towards a General Theory of Bounded Rationality in Mechanisms," *Games and Economic Behavior*, 73 (2011), 360–374.
Che, Y., and D. Hausch, "Cooperative Investments and the Value of Contracting," *American Economic Review*, 89 (1999), 125–147.
Cho, I., "A Refinement of Sequential Equilibrium," *Econometrica*, 55 (1987), 1367–1389.
Cho, I., and D. Kreps, "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102 (1987), 179–221.
Chung, K., and J. Ely, "Implementation with Near-Complete Information," *Econometrica*, 71 (2003), 857–871.
Cremer, J., and R. McLean, "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica*, 56 (1988), 1247–1257.
Dasgupta, P., P. Hammond, and E. Maskin, "The Implementation of Social Choice Rules: Some General Results on Incentive Compatibility," *Review of Economic Studies*, 46 (1979), 185–216.
Fudenberg, D., D. Kreps, and D. Levine, "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory*, 44 (1988), 354–380.
Fudenberg, D., D. Levine, and E. Maskin *Balanced-Budget Mechanisms for Adverse Selection Problems*. (Mimeo, 1991).
Fudenberg, D., and J. Tirole *Game Theory*. (MIT Press, 1991a).
———, "Perfect Bayesian and Sequential Equilibrium," *Journal of Economic Theory*, 53 (1991b), 236–260.
———, "Perfect Bayesian and Sequential Equilibrium," *Journal of Economic Theory*, 53 (1991b), 236–260.
Grossman, S., and O. Hart, "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, 94 (1986), 691–719.
Hart, O., and J. Moore *Some (Crude) Foundations for Incomplete Contracts*. (Mimeo, 2003).
Hendon, E., H. Jacobsen, and B. Sloth, "The One-Shot Deviation Principle for Sequential Rationality," *Games and Economic Behavior*, 12 (1996), 274–282.
Jackson, M., "A Crash Course in Implementation Theory," *Social Choice and Welfare*, 18 (2001), 655–708.
Johnson, S., J. Johnson, and R. Zeckhauser, "Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case," *Econometrica*, 58 (1990), 873–900.
Kreps, D., and R. Wilson, "Sequential Equilibria," *Econometrica*, 50 (1982), 863–894.
Kunimoto, T., and O. Tercieux *Implementation with Near-Complete Information: The Case of Subgame Perfection*. (Mimeo, 2009).
Maskin, E., "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies*, 66 (1999), 23–38.
Maskin, E., and J. Tirole, "Unforeseen Contingencies and Incomplete Contracts," *Review of Economic Studies*, 66 (1999a), 83–114.
———, "Two Remarks on the Property-Rights Literature," *Review of Economic Studies*, 66 (1999b), 139–149.
Matsushima, H., "A New Approach to the Implementation Problem," *Journal of Economic Theory*, 45 (1988), 128–144.
Monderer, D., and D. Samet, "Approximating Common Knowledge with Common Beliefs," *Games and Economic Behavior*, 1 (1989), 170–190.
Moore, J., and R. Repullo, "Subgame Perfect Implementation," *Econometrica*, 56 (1988), 1191–1220.
———, "Nash Implementation: A Full Characterization," *Econometrica*, 58 (1990), 1083–1099.
Muller, E., and M. Satterthwaite, "The Equivalence of Strong Positive association and Strategy-Proofness," *Journal of Economic Theory*, 14 (1977), 412–418.
Myerson, R., "Two-Person Bargaining Problems with Incomplete Information," *Econometrica*, 52 (1984), 461–487.

Saijo, T., "Strategy Space Reduction in Maskin's Theorem: Sufficient Conditions for Nash Implementation," *Econometrica*, 56 (1988), 693–700.
Schmitz, P., "On the Interplay of Hidden Action and Hidden Information in Simple Bilateral Trading Problems," *Journal of Economic Theory*, 103 (2002), 444–460.
Takahashi, S., and O. Tercieux *Robust Equilibria in Sequential Games under Almost Common Certainty of Payoffs*. (Mimeo, 2011).
van Damme, E., and S. Hurkens, "Games with Imperfectly Observable Commitment," *Games and Economic Behavior*, 21 (1997), 282–308.

This page intentionally left blank