Interim Bayesian Persuasion: First Steps*

Eduardo Perez-Richet†

December 30, 2013

Abstract

This paper makes a first attempt at building a theory of interim Bayesian persuasion. I work in a model where a low or high type sender seeks validation from a receiver who is willing to validate high types exclusively. After learning her type, the sender chooses a complete conditional information structure for the receiver. I suggest a solution to this game that takes into account the signaling potential of the sender's choice.

1 Introduction

A new brand of information transmission models views the sender as an *information structure designer* as in Kamenica and Gentzkow (2011). This view builds a (new) connection between the mechanism design literature and the information transmission literature. It departs from traditional models of information transmission (Crawford and Sobel, 1982; Grossman, 1981; Milgrom, 1981; Spence, 1973) since it assumes that the sender commits to an information system before learning her type so that her action is uninformative. In parallel, there is a small but active tradition in mechanism design (Myerson, 1983; Maskin and Tirole, 1990, 1992;

*I thank Pierre Fleckinger, Jeanne Hagenbach, Emir Kamenica, Navin Kartik, Frederic Koessler, Delphine Prady and Joel Sobel for discussions and ideas related to this topic.

†École Polytechnique, e-mail: eduardo.perez@polytechnique.edu

Mylovanov and Troeger, 2012) that considers the possibility that a choice of mechanism by an *informed principal* may be used as a signal by the participants in the mechanism. This possibility radically changes the mechanism design problem, as the chosen mechanism should now be the equilibrium of a signaling game. This short paper proposes a modest first step towards bringing these two approaches together.

I work in a simple framework where a sender, whose type is either high or low, seeks validation from a receiver who wants to validate high types exclusively. The sender chooses an *information system*, that is a space of signals and a conditional distribution of signals for each of her types, from a feasible set. Allowing for potential restrictions on the set of feasible information systems is important to make Bayesian persuasion models more flexible. I consider perfect Bayesian equilibria of the game in which first, the sender chooses an information system after learning her type, and second, the receiver chooses whether to validate. I show that there is no loss of generality in considering only pooling equilibria by an argument partly reminiscent of the *inscrutability principle* of Myerson (1983). In general, perfect Bayesian equilibrium has little predictive power and it is unsatisfying to stop the analysis at this point. My main result is to show that three different refinement concepts lead to the selection of the high type optimal equilibria. The first refinement is a version of the notion of *undefeated equilibria* (Mailath et al., 1993; Umbhauer, 1994) which was developed in the context of signaling games. The second refinement is an adaptation of the notion of *core mechanism* of Myerson (1983), developed for the analysis of informed principal problems, to my framework. Finally, the third refinement builds on the idea of neologism proofness developed in the context of cheap talk games[1] by Farrell (1993). I then analyze the prediction of this method for several common restrictions on feasible information systems.

The idea of this paper is inspired from Perez-Richet and Prady (2012), which shares a similar structure, and where the information system is actually chosen by the receiver in a restricted set, but effectively controlled by the sender insofar as she can choose a complexity parameter that influences the cost of information systems for the receiver. This complexity

[1]See also Mylovanov and Troeger (2012) in which this concept is adapted to informed principal problems.

parameter can then be used as a signal about the type of the receiver. A few other papers analyze the signaling effect of a choice of information transmission technology. In Gill and Sgroi (2012), a monopolist chooses the toughness (a one dimensional parameter) of a pass or fail test for her product, and analyzes the signaling effect of this choice. As in this paper and Perez-Richet and Prady (2012), all equilibria are pooling, and they select undefeated equilibria. In Miyamoto (2013), the information of the sender is two-dimensional, and her choice of an information system on the first dimension can signal something about the second dimension. All these papers parameterize the set of feasible information systems along one dimension, and also make the signaling action of the sender one-dimensional. By contrast, this paper shows how to handle a much richer set of feasible information systems, and by the same token a much richer set of signaling actions.

2 The Model

The Players. A sender of type $t \in \{L, H\}$ seeks validation from a receiver. The receiver is in favor of validation if and only if $t = H$ and believes that type H occurs with probability $p \in (0, 1)$. Without loss of generality, I set the gain from validation to be 1 for both types of the sender. I normalize the payoff of the receiver to be 0 when she makes the right decision, while undue validation entails a loss $\omega_v > 0$ and undue rejection a loss $\omega_r > 0$. If $(1-p)\omega_v < p\omega_r$, the receiver is *(ex ante) pro validation* and incentives are (ex ante) aligned, and if $(1-p)\omega_v > p\omega_r$, she is *(ex ante) pro rejection* and there is a conflict of interests.

Information Transmission. The sender can selectively reveal information to the receiver. She does so by choosing an *information system*, which consists of a realization space $\Sigma \subseteq \mathbb{R}$ and a pair of distribution functions $\pi_L(\cdot)$ and $\pi_H(\cdot)$ in $\Delta(\Sigma)$. This choice may be constrained, so I will denote by $\mathcal{S}$ the set of feasible information systems. For example, in the usual persuasion framework à la Milgrom (1981), the players either reveal their type or reveal nothing, so $\mathcal{S}$ can be described as information systems where $\Sigma = \{0, 1\}$ and the two pairs of distributions available are (π_L, π_H) and $(\tilde{\pi}_L, \tilde{\pi}_H)$, with $\pi_L(0) = \pi_H(1) = 1$, and $\tilde{\pi}_L(0) = \tilde{\pi}_H(0)$. In the

following, I denote by $\mathcal{S}^{rev}$ any set of feasible information systems that restrains the choice of the sender to reveal her type or nothing. In the Bayesian persuasion model of Kamenica and Gentzkow (2011), all information systems are available.[2] Let $\bar{\mathcal{S}}$ denote the corresponding feasible set. I say that an information system (π, Σ) is *fully revealing* whenever π_L and π_H have disjoint support. In $\bar{\mathcal{S}}$ or $\mathcal{S}^{rev}$, full revelation is feasible.

Timing. To allow for comparisons, I consider two possible timings. Under the *ex ante timing*, nature first draws the type of the sender, second the sender chooses a feasible information system, third a signal is generated according to the information system and observed by the players, and finally the receiver decides whether to validate. Under the *interim timing*, the sender learns her type before the second stage.

Equilibrium. I consider perfect Bayesian equilibria. Under the ex ante timing, it just means that the receiver uses the informational content of the signal generated by the information system as a Bayesian. Under the interim timing, the receiver also uses the information contained in the equilibrium choice of strategies by the different types of the sender. I also restrict attention to pure strategies. In a perfect Bayesian equilibrium under the interim timing, the receiver must update her information consistently with equilibrium strategies and with the informational content of the signal generated by the chosen information system. To avoid any paradoxes, I assume that the signal generated by the information system has preeminence off the equilibrium path in the following sense. Suppose that an information system that is not supposed to be part of the equilibrium is chosen and that it generates a signal that could have only come from one of the two types, say t. Then I assume that following this signal, the receiver must believe that she is facing type t with probability 1, even though the equilibrium may have dictated a belief that put probability 0 on t after observing the off path choice of information system. This treatment is consistent with the way evidence is treated in models with hard information.

[2] In fact they constrain Σ to be finite but this is without loss of generality as far as feasible outcomes are concerned. As they show, one could even constrain Σ to be of cardinality 2.

3 Analysis

3.1 Benchmarking

First consider the case $\mathcal{S}^{rev}$ in which the sender is constrained to reveal her type or nothing. Then, under the interim timing, the unique perfect Bayesian equilibrium is fully revealing when the receiver is pro rejection, whereas both full revelation and no revelation are possible equilibrium outcomes in the pro validation case. Under the ex ante timing, full revelation is no longer an equilibrium outcome with a pro validation receiver. If the receiver is pro validation, then no information is revealed, whereas all information is revealed if she is pro rejection.

With a feasible set $\bar{\mathcal{S}}$ and under the ex ante timing, we obtain the framework of Kamenica and Gentzkow (2011). The outcome is a weakening of the results in the $\mathcal{S}^{rev}$ setting with the ex ante timing: when the receiver is pro rejection, not all information is revealed so that the low type is validated with positive probability. In the pro validation case, both types are validated. Some information may be revealed, but in a way that never makes the posterior of the receiver fall below her validation threshold.

3.2 General Results under the Interim Timing

For the players, the relevant properties of an equilibrium are perfectly described by the probability of justified rejection ρ and the probability of justified validation ν. Suppose that a certain information system (π, Σ) is chosen by the sender which leads the receiver to believe that the sender is of the high type with probability β. Then the optimal policy of the receiver is to validate if the realized signal σ is in the set $\Sigma^+ = \{\sigma : \beta\pi_H(\sigma)\omega_r > (1-\beta)\pi_L(\sigma)\omega_v\}$, to reject if σ is in $\Sigma^- = \{\sigma : \beta\pi_H(\sigma)\omega_r < (1-\beta)\pi_L(\sigma)\omega_v\}$, and to randomize in any way if σ is in $\Sigma^0 = \Sigma \setminus \left(\Sigma^+ \cup \Sigma^-\right)$. Consider an optimal policy of the receiver such that she validates with probability λ in case of equality. Then

$$\nu(\beta, \pi, \Sigma, \lambda) = \sum_{\sigma \in \Sigma^+} \pi_H(\sigma) + \lambda \sum_{\sigma \in \Sigma^0} \pi_H(\sigma)$$

and

$$\rho(\beta, \pi, \Sigma, \lambda) = \sum_{\sigma \in \Sigma^-} \pi_L(\sigma) + (1 - \lambda) \sum_{\sigma \in \Sigma^0} \pi_L(\sigma).$$

Then the set of outcomes that can be attained by the sender if her actions lead to a belief β is described by $\mathcal{P}(\beta, \mathcal{S}) = \big\{\big(\rho(\beta, \pi, \Sigma, \lambda), \nu(\beta, \pi, \Sigma, \lambda)\big) : 0 \leq \lambda \leq 1, (\pi, \Sigma) \in \mathcal{S}\big\}$. All equilibrium outcomes must lie in $\bigcup_{\beta \in [0,1]} \mathcal{P}(\beta, \mathcal{S})$, but in fact one can restrain attention to a smaller set.

Proposition 1. *All equilibrium outcomes lie in* $\mathcal{P}(p, \mathcal{S})$.

This is due to the fact that one can restrain attention to pooling equilibria. Suppose first that a fully revealing information system is feasible, and that there exists a separating equilibrium. Then the outcome of the separating equilibrium can be obtained as the outcome of a pooling equilibrium in which both types of the sender pool on a fully revealing information system. Suppose instead that a fully revealing information system is not available. Then the low type can always do better by pooling with the high type, so separation cannot occur in equilibrium. The first part of the argument is reminiscent of the one justifying the *inscrutability principle* of Myerson (1983) for informed principal problems, the second one is of a different nature and relies more on the particular incentive structure of the game.

If perfect revelation is available, then perfect Bayesian equilibria have some predictive power. Indeed, any pooling equilibrium must achieve validation with probability one for the high type, for otherwise a high type sender would deviate to perfect revelation. Therefore an information system (π, Σ) is an equilibrium if and only if it maximizes the utility of the high type under the constraint that the receiver chooses an optimal validation policy given an initial belief equal to the prior, and the choice of information system. Let $\mathcal{H}(p, \mathcal{S}) = \{(\rho, \nu) \in \mathcal{P}(p, \mathcal{S}) : \nu \geq \nu', \ \forall (\rho', \nu') \in \mathcal{P}(p, \mathcal{S})\}$ denote the set of high type optimal equilibrium outcomes, which I assume to be non empty.[3]

Proposition 2. *If a perfectly revealing information system is available, then the set of equilibrium outcomes is exactly the set of high type optimal outcomes* $\mathcal{H}(p, \mathcal{S})$*, and hence* $\nu = 1$ *in any such outcome.*

[3]This set may be empty if, for example, $\mathcal{P}(p, \mathcal{S})$ is an open set of $[0, 1]^2$.

If perfect revelation is not available, however, the predictive power of perfect Bayesian equilibrium is weak, as any feasible information system, and therefore any outcome in $\mathcal{P}(p, \mathcal{S})$, can be supported as a pooling equilibrium if the receiver attributes any alternative choice of an information system to the low type.[4] Furthermore, some of these equilibria seem unreasonable. In this case, equilibrium refinements appear indispensable. My main result is to show that three different but related refinements lead to the selection of the high type optimal outcomes $\mathcal{H}(p, \mathcal{S})$. To gain space, I define these notion more precisely and prove the result in the appendix. The first of these refinements, $\mathcal{R}_1$ is a variant of the notion of undefeated equilibrium (Mailath et al., 1993; Umbhauer, 1994). The concept of undefeated equilibrium relies on the idea that a deviation of the sender from her equilibrium action (π, Σ) to an information system (π', Σ') should be interpreted as an attempt to indicate that she would prefer to play the equilibrium outcome associated to (π', Σ') in $\mathcal{P}(p, \mathcal{S})$. The beliefs of the receiver associated to such a deviation should therefore anticipate the fact that the sender is of a type that benefits from the new equilibrium. If they do not, the original equilibrium is said to be defeated by the new one. The second notion, $\mathcal{R}_2$, is an adaptation of the notion of core mechanisms of Myerson (1983) to my framework so as to take into account the fact that the designer chooses an information system rather than a full mechanism. I call this notion core equilibrium. The third notion, $\mathcal{R}_3$, is a development on the notion of neologism proofness of Farrell (1993) in which a deviating player is allowed to make a suggestion to the receiver as to how her deviation should be interpreted. The idea is to use the suggestion to start a chain of possible types making the deviation by answering the questions of who benefits from the suggestion, who benefits from the sender best-responding to the suggestion etc.

Proposition 3. *All three refinements select exactly the equilibrium outcomes in* $\mathcal{H}(p, \mathcal{S})$.

In the rest of the paper I analyze the consequences of this result under several common restrictions on the set of available information systems. Before doing that, I remark that the

[4]The intuitive criterion does not refine prediction since all information systems perform equally from the point of view of the sender whenever the receiver forms the belief that she faces the high type with probability one.

D1 criterion[5] of Cho and Kreps (1987) may in some cases lead to a different selection, as I show through an example in the appendix.

4 Applications

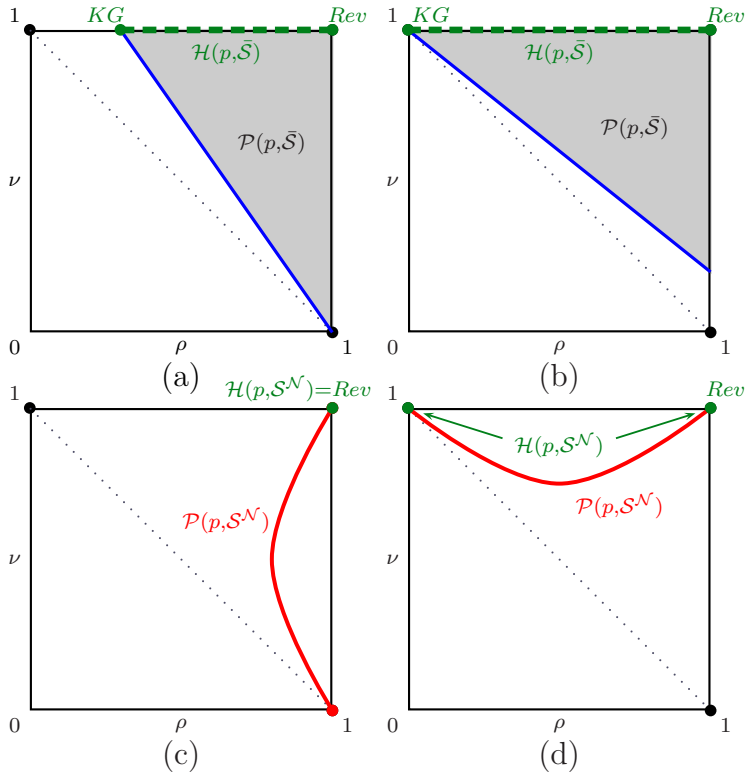

Figure 1: Applications – (a) all information systems available; pro rejection receiver (b) all information systems available; pro validation receiver (c) information systems: one of the normal families; pro rejection receiver (d) information systems: one of the normal families; pro validation receiver.

First, I consider the case where all information systems are available. It is represented in panels (a) and (b) of Figure 1 for, respectively, the pro rejection and pro validation cases. The set of feasible outcomes $\mathcal{P}(p,\bar{\mathcal{S}})$ is the set of policies that lie above the blue line whose equation is given by $\nu p\omega_r + \rho(1-p)\omega_v = \chi$, where the left hand-side is the objective that the receiver tries to maximize, and the right hand-side is the value of this objective function if she always rejects for the pro rejection case and if she always validates for the pro validation case. By

[5]See also Banks and Sobel (1987) and Cho and Sobel (1990).

Proposition 2, the set of equilibrium outcomes is exactly the set of policies such that the high type is validated with probability 1, which is represented by the green dashed line in both panels. To the left of this line, KG denotes the ex ante (Kamenica-Gentzkow) solution. To the right, Rev is the receiver-optimal policy, which can only be attained under full revelation. By refining the solution concept even more, it may be possible to select either of these outcomes in the interim case. If one requires the selected outcome to be Pareto optimal across types, for example, then the ex ante outcome is the unique selection. To select the perfect revelation outcome, one can modify the preferences of the sender as follows. Suppose that the sender cares lexicographically: first, about the probability of validation; second about the belief of the receiver. The prediction under the ex ante timing remains the same as before. Under the interim timing, however, the unique high type optimal outcome now corresponds to full revelation.

Second, I consider a case in which all information systems have $\Sigma = \mathbb{R}$, and π_H is a normal distribution with mean $\ell > 0$ and variance σ^2, while π_L is a normal distribution with mean $-\ell$ and variance σ^2. Then, there are two natural ways to generate a family of feasible information systems: first, by considering variances from $\sigma = 0$ to ∞, and adding a completely uninformative information system; second by letting ℓ go from 0 to ∞, and adding a perfectly revealing information system. In the discussion that follows, both families are denoted by $\mathcal{S}^{\mathcal{N}}$ as they have the same qualitative properties. Panels (c) and (d) in Figure 1 illustrate the properties of these families for, respectively, the pro rejection and the pro validation cases. The red curve represents the set $\mathcal{P}(p, \mathcal{S}^{\mathcal{N}})$ of feasible outcomes, and $\mathcal{H}(p, \mathcal{S}^{\mathcal{N}})$ corresponds to the green dots. In the pro rejection case, full revelation is the unique equilibrium outcome, whereas in the pro validation case, both the full revelation outcome and the completely uninformative outcome leading to certain validation are possible. Under the ex ante timing, the unique prediction would lie somewhere on the upper half of the red curve for the pro rejection case, and at the uninformative outcome for the pro validation case.

Under the pro validation case, it is interesting to consider the prediction when the precision of available information systems (taken to be ℓ in the first family, and the inverse of the variance

in the second family) is only allowed to move in a given range. Then for such a feasible set $\mathcal{S}$, the set $\mathcal{P}(p, \mathcal{S})$ would correspond to a connected portion of the U-shaped red curve in panel (d). Hence the high type optimal outcome would lie at either extreme point of the curve depending on the range of precisions. This means that a small technical change leading to the availability of slightly more precise tests could lead to a dramatic change of outcome from the least informative test to the most informative test. On the other side of the same coin, loosening regulation as to the minimal precision of a test could lead to a dramatic change from the most informative test to the least informative test.

Appendix

A Three Refinements and a Proof

In this section I define the refinements for general sender-receiver games. The type set is denoted by T, $p(\cdot)$ denotes the prior belief, and for any set $S \subseteq T$, $p(\cdot|S)$ denotes the restriction of the prior to S. Then I prove Proposition 3.

$\mathcal{R}_1$: Undefeated Equilibrium. The idea is that an out-of-equilibrium action should be interpreted as an attempt by a player to signal that she would prefer to coordinate on another equilibrium in which this action is played. If the beliefs of the players in the original equilibrium do not anticipate this, the original equilibrium should be discarded. This idea was developed independently in Mailath et al. (1993) and Umbhauer (1994) with small differences. Formally, in a sender-receiver game with general action set for the sender, if T is the type set of the sender and e is the original equilibrium, consider an equilibrium e' and an action of the sender a' which is on the equilibrium path of e' but off the equilibrium path of e'. Then let $T^+ \subseteq T$ be the set of types that strictly prefer e' to e and use action a' in e', and $T^0 \subseteq T$ be the set of types who are indifferent between e and e' and use a' in e'. Unlike Mailath et al. (1993), but like Umbhauer (1994), we do not require all types who use a' in e' to prefer e' to e. Then the belief of the receiver that follows the use of a' in the initial equilibrium e should be in the

convex hull of $p(.|T^+)$ and $p(.|T^+ \cup T^0)$. This means that the receiver should believe that all types in T^+ send a' while types in T^0 may send a' with positive probability. If this does not hold, then e is defeated by e'. The refinement retains only undefeated equilibria.

$\mathcal{R}_2$: Core Equilibria. This refinement is inspired from Myerson (1983). The difference is that the sender announces an information structure instead of a mechanism. When the sender announces a mechanism, his announcement includes the suggestion of a course of action for the receiver, and it is natural to restrict potentially destabilizing mechanisms to be incentive compatible given the beliefs they may generate. When the sender merely announces an information structure, the receiver should best respond given her beliefs, but that does not entail any natural restriction on the announcements of the sender. In order to define core outcomes, I consider a sender-receiver game, with general finite type set T for the sender, and where the sender announces an information system (π, Σ) (defined for the general type set T). Then an information system (π, Σ) is a core information system if it is an equilibrium and there is no other information system $(\pi', \Sigma') \neq (\pi, \Sigma)$ and set $S \subseteq T$, such that for every belief $p(\cdot|S')$ of the receiver, where $S \subseteq S' \subseteq T$, any type $t \in S$ strictly prefers the outcome obtained when the receiver best responds to (π', Σ') to the initial equilibrium outcome. The motivation is as follows. Suppose that (π, Σ) is not a core information system. Then there exists a subset of types S that would benefit from any beliefs that restricts the prior to any superset of S. Then any type in S could credibly announce (π', Σ'), and tell the receiver "my type is in S." The receiver does not have to believe that the sender is indeed in S, but she should account for the fact that all types in S are strictly better off as long as she believes that they make this statement.

$\mathcal{R}_3$: A Variation on Neologism Proofness. As in the former paragraph. I describe the refinement for a sender-receiver game with general type set T, and where the sender announces an information system (π, Σ). As in Farrell (1993), I assume that statements of the kind "my type is in S" are available for every $S \subseteq T$. Consider an equilibrium e and an information system (π', Σ') which is never played in e. When deviating to (π', Σ'), the sender can also

announce that her type belongs to some set $S_0 \subseteq T$. Then let S_1 be the set of types that strictly benefit from the best response of the receiver to (π', Σ') under the belief $p(\cdot|S_0)$ relative to the initial equilibrium, and so on, so that S_{k+1} is the set of types that strictly benefit from the best response of the receiver to (π', Σ') under the belief $p(\cdot|S_k)$ relative to the initial equilibrium. The sequence stops if the empty set is ever reached. The types in $\bigcup_k S_k$ are those who could be tempted to use the deviation (π', Σ') together with the announcement "my type is in S_0." Therefore the initial equilibrium is deemed unreasonable if it can only be supported by a belief $q \in \Delta(T)$ that does not lie in the convex hull of the set $\{p(\cdot|S_1), p(\cdot|S_2), p(\cdot|S_3), \cdots\}$. If the sequence is empty ($S_1 = \emptyset$), then all beliefs are allowed. Note the difference with Farrell (1993), which would require the existence of a set S_0 such that $S_1 = S_0$ (in (Farrell, 1993), the deviation is the announcement itself, whereas here it consists in a choice of a different information system accompanied with the announcement).

Proof of Proposition 3. First consider $\mathcal{R}_1$. Let (π, Σ) be an information system such that the associated equilibrium outcome in $\mathcal{P}(p, \mathcal{S})$ is not in $\mathcal{H}(p, \mathcal{S})$ (so full revelation must not be available).Then consider any information system $(\pi', \Sigma') \neq (\pi, \Sigma)$ such that the associated equilibrium outcome is in $\mathcal{H}(p, \mathcal{S})$. Then the high type must prefer the new outcome to the original. Suppose first that the low type prefers the original equilibrium outcome. Then after observing the deviation (π', Σ'), the receiver who, according to $\mathcal{R}_1$, is assumed to interpret it as an attempt to coordinate on the new equilibrium must believe that this message comes from the high type. If she did, however, the original equilibrium would not be an equilibrium as both types would benefit by deviating to (π', Σ'). Suppose now that the low type is weakly prefers the new equilibrium. Then after observing the deviation (π', Σ'), the receiver must believe that she faces the high type with probability $p' \geq p$. However the original equilibrium cannot be supported by such a belief, since by deviating to (π', Σ') the high type would get both a more favorable belief p' and a more favorable information system. This shows that all selected equilibrium outcomes lie in $\mathcal{H}(p, \mathcal{S})$. To show that the two sets are in fact equal, consider an information system (π, Σ) that leads to an equilibrium outcome in $\mathcal{H}(p, \mathcal{S})$. Since the high

type cannot improve her situation, the refinement does not prevent from believing that any deviation is originated by the low type, and such beliefs clearly support the equilibrium.

Now consider $\mathcal{R}_2$. Let (π, Σ) be an information system such that the associated equilibrium outcome in $\mathcal{P}(p, \mathcal{S})$ is not in $\mathcal{H}(p, \mathcal{S})$. Consider another information system (π', Σ') with an associated equilibrium outcome in $\mathcal{H}(p, \mathcal{S})$. Then let $S = \{H\}$. Clearly the high type prefers the outcome associated to the information system (π', Σ') and the belief $p(\cdot|S)$ since the latter must put probability one on the high type. Now consider $S' = \{H, L\}$. Then $p(\cdot|S')$ is simply the prior, and since (π', Σ') is in $\mathcal{H}(p, \mathcal{S})$, the best response to (π', Σ') when the belief is the prior leads to a better outcome than the equilibrium associated with (π, Σ). Therefore the initial equilibrium is not a core equilibrium. This proves that all core equilibrium outcomes lie in $\mathcal{H}(p, \mathcal{S})$. Now consider an information system (π, Σ) such that the associated equilibrium outcome lies in $\mathcal{H}(p, \mathcal{S})$. It can be supported by the belief that any other choice is due to the low type. Suppose that this equilibrium is not a core equilibrium. Then the high type would have to strictly prefer the outcome associated with a different information system (π', Σ') under the belief $p(.|T)$, which is simply the prior. But then that would contradict the fact that (π, Σ) together with the prior leads to a high type optimal outcome.

Now consider $\mathcal{R}_3$. Let (π, Σ) be an information system such that the associated equilibrium outcome in $\mathcal{P}(p, \mathcal{S})$ is not in $\mathcal{H}(p, \mathcal{S})$. Consider another information system (π', Σ') with an associated equilibrium outcome in $\mathcal{H}(p, \mathcal{S})$. Suppose that the receiver deviates from the original equilibrium by choosing (π', Σ') and at the same time suggests to the receiver that she is the high type, so $S_0 = \{H\}$. The receiver must realize that both types would benefit if she were to believe the suggestion, so $S_1 = \{H, L\}$, and the corresponding belief is exactly the prior. If that is indeed the receiver's belief, she will reproduce the outcome associated with (π', Σ'). This outcome makes the high type strictly better of. There are two cases. First, if it is does not make the low type strictly better of, then $S_2 = \{H\}$, and the sequence generated is therefore $S_k = \{H\}$ for every even k, and $S_k = \{L, H\}$ for every odd k. Second, if both types are better of under the outcome obtained when the receiver best responds to (π', Σ') with a belief equal to the prior. Then $S_2 = S_1 = \{L, H\}$, and that pins down the sequence $S_k = \{L, H\}$ for

every $k \geq 1$. In both cases, the possible beliefs that support the initial equilibrium following a deviation to (π', Σ') must lie in $[p, 1]$, but that clearly makes this deviation profitable for the high type, so the initial equilibrium cannot satisfy the refinement. Note that we could have used the suggestion $S_0 = \{L, H\}$ to get the same result. This proves that all equilibrium outcomes that satisfy $\mathcal{R}_3$ lie in $\mathcal{H}(p, \mathcal{S})$. Now consider an information system (π, Σ) such that the associated equilibrium outcome lies in $\mathcal{H}(p, \mathcal{S})$. Consider any deviation (π', Σ'). If the suggestion of the receiver is $\{L\}$, then $S_1 = \emptyset$ and the belief that puts all the weight on the low type is allowed following this deviation and it supports the original equilibrium. If the suggestion is $\{L, H\}$, then the associated belief is the prior, but since the original equilibrium is high type optimal, the set S_1 is either $\{L\}$ or the empty set. If $S_1 = L$, then $S_2 = \emptyset$. In both cases, the belief that puts all the weight on the low type is allowed and supports the original equilibrium. Finally, suppose the suggestion of the receiver is $S_0 = \{H\}$. Then $S_1 = \{L, H\}$, and S_2 is either $\{L\}$ or the empty set. So the prior is a possible belief in both cases, and it supports the original equilibrium. □

B D1: An Example

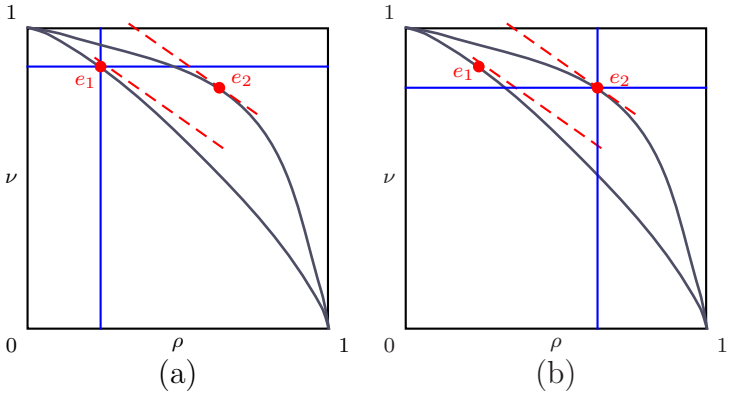

Figure 2: Example – The high type optimal equilibrium is e_1, but the only equilibrium that satisfies D1 is e_2. (a) shows why e_1 cannot satisfy D1, while (b) shows why e_2 satisfies D1.

In this example the sender has only two information systems I_1 and I_2 available, and the sets of possible outcomes generated by these information systems under all beliefs, $\cup_{\beta \in [0,1]} \mathcal{P}(\beta, \{I_1\})$

and $\cup_{\beta\in[0,1]}\mathcal{P}(\beta,\{I_2\})$, are represented respectively by the lower and the higher curve on Figure 2. The receiver is pro validation, and her indifference sets over outcomes are represented by the dashed red lines. So the two possible equilibria are e_1 and e_2. Consider e_1 and a deviation in which the sender announces I_2 instead of I_1. The best responses of the receiver that are preferred by the high type given this deviation are the ones that lie on the portion of the higher curve that is above the horizontal blue line in panel (a), while those that are preferred by the low type are the ones that lie on the portion of the higher curve that is to the left of the vertical blue line in panel (a). Clearly, according to D1, the receiver should attribute the deviation to the high type, but e_1 cannot be supported if that is the case. Now consider e_2 and a deviation in which the sender announces I_1. The best responses of the receiver that are preferred by the high type given this deviation are the ones that lie on the portion of the lower curve that is above the horizontal blue line in panel (b), while those that are preferred by the low type are the ones that lie on the portion of the lower curve that is to the left of the vertical blue line in panel (b). According to D1, the receiver should attribute the deviation to the low type, and since this belief is compatible e_2, this equilibrium passed the test. However it is easy to see that e_1 is the unique high type optimal equilibrium.

References

Banks, J. S. and J. Sobel (1987): "Equilibrium Selection in Signaling Games," *Econometrica*, 55, 647–661.

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

Cho, I.-K. and J. Sobel (1990): "Strategic Stability and Uniqueness in Signaling Games," *Journal of Economic Theory*, 50, 381–413.

Crawford, V. and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–1451.

Farrell, J. (1993): "Meaning and Credibility in Cheap Talk Games," *Games and Economic Behavior*, 5, 514–531.

Gill, D. and D. Sgroi (2012): "The Optimal Choice of Pre-Launch Reviewer," *Journal of Economic Theory*, 147, 1247–1260.

Grossman, S. J. (1981): "The Informational Role of Warranties and Private Disclosure about Product Quality," *Journal of Law and Economics*, 24, 461–483.

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

Mailath, G. J., M. Okuno-Fujiwara, and A. Postlewaite (1993): "Belief-Based Refinements in Signaling Games," *Journal of Economic Theory*, 60, 241–276.

Maskin, E. and J. Tirole (1990): "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values," *Econometrica*, 58, 379–409.

——— (1992): "The Principal-Agent Relationship with an Informed Principal, II: Common Values," *Econometrica*, 60, 1–42.

Milgrom, P. (1981): "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics*, 12, 380–391.

Miyamoto, S. (2013): "Signaling by Blurring," Working Paper.

Myerson, R. B. (1983): "Mechanism Design by an Informed Principal," *Econometrica*, 51, 1767–1797.

Mylovanov, T. and T. Troeger (2012): "Informed Principal Problems in Generalized Private Value Environments," *Theoretical Economics*, 7, 465–488.

Perez-Richet, E. and D. Prady (2012): "Complicating to Persuade?" Working Paper.

Spence, M. (1973): "Job Market Signaling," *Quarterly Journal of Economics*, 87, 355–374.

Umbhauer, G. (1994): "Forward Induction, Consistency, Preplay Communication and Epsilon Perturbations," Mimeo Beta, Univesité Louis Pasteur, Strasbourg, France.