# Belief Updating and Equilibrium Refinements in Signaling Games*

Graciela Rodríguez Mariné
University of California, Los Angeles
Department of Economics
November, 1995

**Abstract**

This paper explores the consequences that alternative ways of drawing inferences from deviations have upon the refinements of sequential equilibrium in signaling games. It addresses, in the first place, the following question: *what kind of inference, regarding the type of sender, could the receiver draw when deviations, or off-the-equilibrium path signals, are thought to be meaningless?* Our answer is that there is nothing the receiver can learn in this case, and we therefore suggest that he use the prior distribution over the types as his beliefs after the deviation. This criterion may refine the set of sequential equilibria but equilibrium need no longer exist. Second, the paper analyzes the extent to which the *Intuitive Criterion* and *Divinity* refine the set of sequential equilibria. Both refinements assume that off-the-equilibrium path signals could be intentional. However, this approach has two caveats: on the one hand, not every game can consistently encompass the compatibility of common knowledge of rationality and fully intentional off-the-equilibrium path behavior (for instance, when the equilibrium payoff for every sender dominates any alternative payoff. In this case, deviations can not be taken as signals). On the other hand, in some games, the sender may benefit from a deviation depending not on his type but on the answer of the receiver. This is another situation in which a deviation should not be taken as a signal. In both cases, *Divinity* does not refine the set of sequential equilibria and we suggest that the prior distribution over the types be used as the posterior. The set of *Divine* equilibria could in this way be narrowed down.

---

*This paper is a revised version of the second chapter of the doctoral dissertation submitted to UCLA in November, 1995.

## 1 Introduction

Many issues modeled in economics are concerned with non-cooperative scenarios in which there is one agent who is privately informed about the state of nature and another who is uninformed. The informed agent chooses an action which is observed by the uninformed agent who, after drawing certain inferences, chooses an action in response. The payoffs to both parties depend on the state of nature and the actions taken. A signaling game is a stylization of this type of scenario: by taking an action, the informed agent sends a signal or message to the uninformed who, based upon this observation, constructs beliefs regarding the true state of nature which is unknown to him. The uninformed player responds by taking an action which is a best response to these beliefs. The message sent by the informed player maximizes his expected payoff given his beliefs concerning the response of the uninformed player.

A Nash equilibrium of this type of game is a profile of behavioral strategies such that each of them constitutes a best reply with respect to the other. This means that no player can individually gain by deviating from a Nash equilibrium when the other player plays in accordance with it. However, in order to decide how to play, the informed party needs to deliberate about player 2' responses to his different messages. The only requirement that Nash equilibrium imposes is that the equilibrium response by player 2 to unsent messages deters the informed party from abandoning the strategy prescribed by the equilibrium.

The concept of Nash equilibrium requires that each player chooses a strategy which maximizes his expected payoff, assuming that the other players play *in accordance* with the equilibrium. For this reason, it *may* prescribe responses that are not optimal in the face of a deviation. When this occurs, a player may profit from playing a non-equilibrium strategy in response to a deviation. It is possible that the player who first deviated did so expecting or believing that his deviation would trigger further deviations. In other words, this equilibrium is not *self-enforcing*.

It is accepted that, in order to be self-enforcing in the sense just described, an equilibrium must recommend maximizing behavior at *every* possible situation at which a player may have the chance to play. In the case of games of perfect information, the concept of subgame perfection is sufficient to fulfill this requirement. However, this does not necessarily hold in the case of games of imperfect information; namely, when there is at least one move by a player who does not know which action preceded his turn. Among the solutions proposed to guarantee maximizing behavior at all nodes, while encompassing scenarios in which a player may not know with certainty which action was played before, there is the notion of sequential equilibrium.

A Nash equilibrium is sequential *if* there exists a probability distribution over the states of nature such that the uninformed agent maximizes his expected payoff in the face of a deviation. In other words, under the requirements of sequential equilibrium, responses to *unsent* messages should *also* be best replies, based upon consistent beliefs. Along the equilibrium path, players use Bayes' rule to compute their beliefs regarding the states of nature. By this rule, the uninformed party can calculate the likelihood of every state of nature, conditional on every message that is sent under equilibrium. However, it is assumed that off-the-equilibrium messages constitute zero probability events. For this reason, beliefs based upon off-the-equilibrium messages are left undetermined, given that Bayes' rule is not defined for conditioning events of this sort.

The notion of sequential equilibrium only imposes an *existence* requirement upon the set of beliefs that support players' best responses at nodes which are off-the-equilibrium path. These beliefs do not need to satisfy any further requisite, neither is there a rule to compute them. As a consequence, many equilibria may remain, even after requiring optimal responses at every information set.

In signaling games, sequentiality imposes no restrictions upon the interpretation of messages which are sent off-the-equilibrium path. For instance a player may believe that had his node been reached his opponent would have played a strictly dominated strategy. This may happen because beliefs formed after deviations are not constructed upon the assumption that they signal intentions on the part of the informed player and therefore are considerably unrestricted.

With the purpose of further refining the concept of sequential equilibria several equilibrium notions have been introduced in the literature. Among them are Cho and Kreps' *Intuitive Criterion* [3] and Banks and Sobel's *Divinity* and *Universal Divinity* [1]. The aim of these notions is to restrict possible inferences at information sets that are off-the-equilibrium path by eliminating those beliefs which do not survive some stylization of the hypothesis that deviations could be intentional.

Consider an equilibrium under which regardless of the state of nature the informed party always sends the same message; this is called a pooling equilibrium. Given the distribution from which nature is drawing its states, the uninformed player calculates by Bayes' rule the probability that he faces each state, conditional on the equilibrium message he has received. In this case, the conditional posterior of each state, given the equilibrium message, is equal to the prior probability of the state. Therefore, under equilibrium no new information is revealed; the uninformed player follows the prior distribution over the states of nature. Now assume that a deviation occurred. This occurrence does not intrinsically confer any information, unless we define some concept of rationality together with a theory of how to interpret deviations. Assume that players are rational, as it is typically argued in

the literature. The question is how to model this deviation. Applying the framework of possible worlds as it is done in [10], we look for the smallest departure from the equilibrium under analysis that will bring the counterfactual world of a deviation. If players are rational at least two possible interpretations seem feasible: 1) deviations are mere hypothetical constructions; that is, they are not intentional and therefore can not reveal further information and 2) deviations are intentional and should be analyzed as the outcome of a rational decision process.

Consider the first case. If a deviation does not confer a signal, one option for player 2, is to adopt the prior distribution -which is assumed to be common knowledge- as his beliefs regarding the likelihood of the states of nature. Two questions can be posed: i) why should a deviation be thought of as unintentional? And ii) why should the second player use the prior distribution as his beliefs off-the-equilibrium path in this case?

With respect to the first question, we consider that a deviation may be rationally chosen, and therefore intentional, *only if* it is expected to be profitable. However, by definition, deviations from equilibrium may be profitable *only when* they lead to further deviations. In addition, these "further deviations" also need to be compatible with the assumption of rationality and fully intentional behavior. As it was already illustrated in [10] not every game can consistently encompass the compatibility of the assumption of common knowledge of rationality and fully intentional behavior[1]. Therefore, the extent to which a deviation could be intentional, crucially depends on the structure of the game under consideration.

With respect to the second question, the answer can only be solved in an empirical context. A player may think that in the face of unreliable inferences he may trust a more genuine piece of information; namely, the prior distribution over types. Alternatively, the player may ask himself why he should revise his prior probabilities, which are common knowledge, conditional on the occurrence of an event that is not supposed to confer information. However, there is no unique way to address this issue in a counterfactual scenario, which is by nature, ambiguous or context dependent.

Consider now the case in which deviations are intentional and players rational. Both the *Intuitive Criterion* and *Divinity* propose a framework to restrict beliefs off-the-equilibrium path in situations in which deviations can be rationally explained.

---

[1] Consider a typical three-legged centipede game: it can not be common knowledge that the first player intentionally deviates at the first node by leaving the money with the intention of taking it at the last node. If the second player believes that this is his intention, then he would take the money at the second node rendering the intention unrealizable.

Roughly speaking, Cho and Kreps' *Intuitive Criterion* proposes that states of nature which can not lead to a profitable deviation by the informed player be eliminated from the game or attached zero probability by the uninformed player. In addition, Banks and Sobel assume that the posterior conditional probabilities of those states of nature which, relative to others, are less likely to lead to an intentional deviation be revised downwards with respect to their prior unconditional probabilities. Based upon this refinement, they further propose an iterative procedure, whose equilibrium outcome they call *Divine*.

Both the *Intuitive Criterion* and *Divinity* analyze the intentionality of deviating play while taking for granted that equilibrium responses would have occurred in response to equilibrium play. Although consistent with the features of the equilibrium world that one may want to preserve, this assumption has been challenged by Van Damme ([12] page 281) and it is discussed in depth in section 2.

Regarding the *Intuitive Criterion*, it is not clear why players should be completely eliminated from the game given that the sequential reasoning starts with a tentative hypothesis that can lead to further reactions and reconsiderations of the initial assumption. This issue has been pointed out by Van Damme as a drawback of this method ([12] page 281-282). On the other hand, this test does not impose any restriction upon the set of sequential equilibria when all the types could -to some extent- potentially benefit from a deviation.

Regarding *Divinity,* when the equilibrium payoffs for every sender dominate any alternative payoff in the game, *any* conjecture by the receiver supports the equilibrium. In this case, Divinity does not refine the set of sequential equilibria. Banks and Sobel justify this feature of their proposal by explaining that in this case the receiver should be *truly surprised*. However, if this is the case, then a deviation should be considered meaningless and the comment provided in the previous case holds. Moreover, there is another circumstance in which *Divinity* does not refine the set of sequential equilibrium outcomes. This occurs when the issue of whether player 1 may benefit from a deviation, depends only upon the reply by player 2 and not upon his type. A deviation in this case should also be considered as *truly surprising* in Banks and Sobel's terminology. In section 4 we introduce a modification to Divinity based upon the assumption that whenever the message of the informed player does not convey a signal, the uninformed player uses the prior distribution over the states of nature. With this modification, the set of sequential equilibria can be further refined although existence is not longer guaranteed.

The paper is organized as follows. Section 2.1 presents a formalization of a signaling game and introduces the concept of Nash and Sequential equilibria. Section 2.2 introduces the beer-quiche game from Cho and Kreps [3] together with their *Intuitive Criterion*. Section 2.3 presents Banks and Sobel's concept of *divine equilibria* and *universally divine*

*equilibria* together with some examples that illustrate them. Section 2.4 compares the concepts presented in its previous two sections. Section 3 presents an alternative criterion to restrict beliefs at information sets that are off-the-equilibrium path, based upon the interpretation of Lewis' theory of counterfactuals in [10], and characterizes the situations in which equilibrium exists. Section 4 introduces a variation to Banks and Sobel's Divine equilibria and presents an alternative iterative procedure to eliminate implausible sequential equilibria. There is a final section which analyzes the extent to which these refinements are plausible and their relationship to the concept of rationality.

## 2 Signaling Games

Consider the following two players non-simultaneous move game: the first player, called the *sender* (S) chooses, after learning his *type* t, a *message* or *signal* m, from a finite set M(t). This type is drawn from a finite set T according to a probability distribution $\pi$ which is common knowledge among both players. The sender's type, that is the particular realization of $\pi$, is unknown to the second player, the *receiver* (R), whose decision consists in choosing an action, a, from a finite set A(m) after observing the sender's message, $m \in M(t)$. This response finishes the game. The resulting players' payoffs, u(t,m,a) and v(t,m,a) respectively, are determined by the message, the responding action and the sender's type.[2]

The rules of the game represented by $\Psi=(T,M,A,\pi,u,v)$ are common knowledge among the players. The asymmetry consists in the fact that the receiver, that is player 2, does not know a piece of information that player 1 knows; namely player 1's type.

Let P(T), P(M) and P(A) be the set of probability distributions over T, M and A respectively. An element $\tau_m$ of P(T) represents a set of beliefs by player 2 concerning the likelihood of types $t \in T$ after receiving message m. Let $\underline{\tau}=(\tau_m)_m$ denote a system of beliefs of player 2. The elements of P(M) and P(A) will be denoted by $\mu$ and $\alpha$ respectively.

Let $p=(p_t)_t$ and $r=(r_m)_m$ respectively represent the sender's and receiver's behavioral strategies with $p_t \in P(M)$ for all $t \in T$ and $r_m \in P(A)$ for all $m \in M$. Given the sender's type t, p( . ;t) is a probability distribution over M(t) and given the sender's message, r( . ;m) is a probability distribution over A(m).

The expected payoff to type t when he sends the mixed message $\mu$ and player 2 responds with strategy r is: $u(t,\mu,r) := \Sigma_{m,a}\ \mu(m)\, r_m(a)\, u(t,m,a)$.

---

[2]As a simplifying assumption and without loss of generality we shall assume that M (t), the set of messages available to type t, is the same for all types. By the same token we shall also assume that A(m), the set of actions available after message m, is the same for all messages.

On the other hand, the expected payoff of a player who faces message m when he has beliefs $\tau$ and responds with a mixed action $\alpha$ is: $v(\tau,m,\alpha):= \Sigma_{t,a}\ \tau(t)\ \alpha(a)\ v\ (t,m,a)$.

Let p(m) be the probability that message m is chosen under strategy p:

$p(m):=\Sigma_{t \in T(m)}\ \pi(t)\ p_t(m)$.

The beliefs of player 2, who responds to message m, calculated according to Bayes' rule are as follows: $\tau_m^{\,p}(t) := \pi(t)p_t(m) / p(m)$ if $p(m) >0$.

Let the best response by type t against r be

$BR_t\ (r):= \underset{\mu}{\text{argmax}}\ u(t,\mu,r)$,

and let the best response to message m by player 2 if he holds beliefs $\tau$ be

$BR_m\ (\tau):= \underset{\alpha}{\text{argmax}}\ v(\tau,m,\alpha)$.

A Nash equilibrium for the game under consideration consists of a pair of behavioral strategies (p,r), such that

$p_t \in BR_t\ (r)$ for all $t\in T$, and $r_m \in BR_m\ (\tau_m^{\,p})$ for all $m\in M$ with $p(m)>0$.

Under equilibrium, the sender maximizes his expected utility given the receiver's response and, the receiver, after receiving the equilibrium message m, computes by Bayes' rule the probability that, given m, the sender is type t, for every t. In a second step, player 2 maximizes his expected utility using these probability assessments as his beliefs.

### 2.1 Sequential Equilibrium

There is general agreement upon the need of prescribing rational behavior at information sets that are off-the-equilibrium path for the equilibrium to be self-enforcing. The problem that remains, however, is how to model or think about these zero probability events. The concept of sequential equilibria is one of the notions that has been proposed to deal with this problem. Under this concept, players are modeled as expected utility maximizers *at all nodes*. Furthermore, in the face of uncertainty, players are assumed to behave in the following way:

(i) they calculate conditional probabilities by Bayes' rule along the equilibrium path. This guarantees the consistency requirement upon beliefs;

(ii) they form beliefs off-the-equilibrium path by constructing posterior probability distributions on each information set that is not reached under equilibrium;

(iii) at each information set players assume that in the remainder of the game players will play according to the equilibrium under consideration.

A strategy profile (p,r) is a sequential equilibrium if there exists a system of consistent beliefs $\underline{\tau}=(\tau_m)_m$ such that each player's strategy maximizes his expected utility given his probability assessments on *and* off the equilibrium path.

Formally a triple $(p,r,\underline{\tau})$ is a sequential equilibrium if:

$p_t \in BR_t(r)$ for all $t \in T$,

$r_m \in BR_m(\tau_m)$ for all $m \in M$,

$\tau_m = \tau_m{}^p$ for all m with $p(m)>0$.

It is clear that a sequential equilibrium need not be optimal with respect to *all* beliefs. Moreover, the beliefs which support a sequential equilibrium need not be *sensible*. A response to a deviation by the sender may be based upon the belief that the sender has played a dominated strategy. The reason is that deviations are implicitly treated as mistakes and therefore not connected with payoffs. Sequentiality only adds optimality at every node and consistency of beliefs *along* the equilibrium path.

### 2.2 The Intuitive Criterion

Consider the following game taken from Cho and Kreps [3]. Nature selects a type for player 1 who can be either *strong* (s) or *weak* (w). That is, T={s,w}. The probability distribution over the types is given by $\pi = \{\pi(s)=0.9, \pi(w)=0.1\}$. Without knowing whether player 1 is strong or weak and regardless of the message, player 2 has to decide whether to duel player 1 or not; that is A(m)=A={d, ~d}. If player 2 duels a strong player 1 he receives a payoff of 0; if he duels a weak player 1 he receives a payoff of 1. If player 2 decides not to duel his payoffs are 0 if the opponent is weak and 1 if he is strong. In other words player 2 would wish to duel if he believed his opponent was weak and not duel otherwise. After this response, the game ends and the players receive their payoffs. In any event, player 1, who knows his type, has to choose between having beer or quiche for breakfast; that is, M(s)=M(w)={b,q}. Other things equal in terms of player 2's response, the weak type prefers a breakfast of quiche whereas the strong prefers a breakfast of beer. The payoffs to the players are depicted in Figure 1.

| beer | duel | ~duel | quiche | duel | ~duel |
|---|---|---|---|---|---|
| strong | 1,0 | 3,1 | strong | 0,0 | 2,1 |
| weak | 0,1 | 2,0 | weak | 1,1 | 3,0 |

Figure 2

This game has two set of Nash equilibria associated with the following outcomes:

1) player 1 -regardless of his type- has beer for breakfast; player 2 avoids the duel when the message is beer and otherwise duels with a probability of at least 0.5. In terms of our notation: p(b/s) = p(b/w) =1; r(~d/b)=1; r(d/q) ≥0.5 . This is indeed a sequential equilibrium provided that player 2's beliefs are such that $\tau_b^p$ (s)≥0.5 and $\tau_q$ (s)≤ 0.5. The first inequality is guaranteed because along the equilibrium path, player 2's conditional beliefs are equal to the prior distribution (by Bayes' rule $\tau_b^p$ (s) = π(s)=0.9) and this prevents him from dueling. The second inequality is not constrained because there is no rule to compute these beliefs.

Existence of Nash equilibrium requires that player 2 duels if quiche with a probability of at least 0.5, because this response off-the-equilibrium path prevents the deviation by the first player regardless of his type. This Nash equilibrium is also sequential provided that player 2 believes that, had player 1 had a breakfast of quiche, it would have been more likely that he was weak.

2) player 1 -regardless of the type- has quiche for breakfast; player 2 avoids the duel when the message is quiche and otherwise duels with a probability of at least 0.5. In terms of our notation: p(q/s) = p(q/w) =1; r(~d/q)=1; r(d/b) ≥0.5 . This is indeed a sequential equilibrium provided that player 2's beliefs are such that $\tau_q^p$ (s)≥0.5 and $\tau_b$ (s)≤ 0.5. The first inequality is again guaranteed, because by Bayes' rule $\tau_q^p$ (s) = π(s)=0.9 and these beliefs prevent him from dueling. Off the equilibrium path, player 2 duels with a probability of at least 0.5 and this prevents the deviation by the first player regardless of his type. As before, this equilibrium is also sequential when player 2 believes that, had player 1 had beer for breakfast, it would have been more likely that he was weak.

Cho and Kreps argue that the equilibrium outcome in which both types drink quiche is not sensible. As we saw previously, this outcome is sequential because dueling if beer is a best response when player 2 assigns a probability of at least 0.5 to the event that is the weak type, the one who sent that message. Cho and Kreps argue that these beliefs are not sensible because the weak type can not profit from drinking beer relatively to the payoff which he receives under equilibrium if he has quiche. They assert that if this reasoning is common knowledge among the players, it is clear that player 2 should believe that the strong type is the more likely to have deviated and therefore he should not duel. In this circumstance, the strong confirms the profitability of his deviation and this breaks the equilibrium outcome under analysis.

The formalization of this criterion can be described in the following way:

Let T(m) represent the set of types for whom message m is available. For each message m' sent off-the-equilibrium path define S(m') to be the set of types t whose

equilibrium payoffs, u*(t), exceed the best payoff which they can possibly obtain if they deviate. In other words, $t \in S(m')$ iff:

$$u^*(t) > \max_{a \in BR(\tau(T(m')),m')} u(t,m',a)$$

An equilibrium outcome fails the *Intuitive Criterion* if there is some type $t' \in T$ such that:

$$u^*(t') < \min_{a \in BR(\tau(T(m') \setminus S(m')),m')} u(t',m',a)$$

An equilibrium outcome fails the Intuitive Criterion when there is a type who, by deviating can profit relatively to the payoff that he obtains under equilibrium, while facing a best reply to beliefs that exclude types who could never gain by deviating. In the beer-quiche game the strong type obtains a payoff of 2 under equilibrium. However, he can reach a payoff of 3 if he deviates to beer and player 2 does not duel, given his revised expectations that player 1 is the strong type.

However, it has been pointed out both by Cho and Kreps [3] and Van Damme [12] that if this reasoning is taken one step further and still assumed to be common knowledge among the players, one concludes that *if* beer is a sure sign of a strong type *then* quiche is a sure sign of a weak type. By Bayes rule if $\tau(s/b)=1$ then $p(b/w)=0$. In other words, if beer signals the strong type then the weak type must have quiche for breakfast. Does this imply that quiche *signals* the weak type? To reply this question with an affirmative answer we need to assume that the strong type does not have quiche for breakfast which is not equivalent to asserting that beer can only be chosen by a strong type. To imply that quiche signals the weak type we need to assume that the strong type, apart from being rational, believes that, if he has beer the chance that he will face a duel is less than 0.5.

Cho and Kreps disclaim the argument that quiche signals the weak type by asserting that Nash equilibrium "is meant to be a candidate for a mode of self-enforcing behavior that is common knowledge among the players." They conclude that in order to test an equilibrium outcome one should start with the hypothesis that the corresponding equilibrium is common knowledge and then look for contradictions. In any case, the conclusion that the weak type is better off by not having quiche, leads to conclude that the quiche equilibrium outcome is not self-enforcing. However, one could also say that if the reasoning under consideration implies that beer does not signal a strong type then the initial hypothesis that beer signaled a strong type should be rejected instead of the equilibrium outcome itself. Van Damme suggests this type reasoning as a counterargument that beer off-the-equilibrium path signals the strong type. However, he does not offer a way of solving this dilemma.

Cho and Kreps' opinion is valid if one thinks of an equilibrium as a recommendation to the players that *guarantees* them a certain payoff. Imagine that the players are told that if they have quiche for breakfast, then it is certainly the case that the second player will not duel. Now they need to decide whether a deviation can improve the equilibrium payoff which they will receive with certainty. In this case, Cho and Kreps' analysis results appropriate.

One can also think of an equilibrium as a set of consistent propositions within a language based upon a behavioral assumption and a theory of how to analyze deviations. Under this approach, the logical consequences of the propositions that define the equilibrium should also be taken as part of it. From this point of view, an equilibrium is a set of requirements whose consistency needs to be tested. An equilibrium can be disregarded as self-enforcing as long as we find an internal contradiction within this set of requirements. For instance, the original quiche outcome is sequential and becomes inconsistent as soon as we consider the interpretation that beer signals a strong type. Without this further requirement, that introduces constraints upon off-the-equilibrium path beliefs, the quiche equilibrium outcome is internally consistent. Assuming that beer signals a strong type while fixing the equilibrium provides the strong type with incentives to deviate. However, if by assuming that beer signals the strong type, we accept as a logical consequence that quiche signals the weak type, then we give incentives to player 2 and the weak type to deviate from the original equilibrium. Either way, the internal consistency of the equilibrium outcome is lost.

One could also ask the following question: why should players on the one hand, have their equilibrium payoffs guaranteed and on the other, suppose that if they had deviated *another* deviation would have occurred in response? or why should equilibrium responses to deviations be relaxed and not equilibrium responses to equilibrium play? The framework developed in [10] provides the following answer. The factual world is that in which the equilibrium is played. In this world every player conforms to the equilibrium. Worlds in which deviations occur constitute counterfactual scenarios where many things are possible. In order to reach a counterfactual scenario, one needs to drop at least one feature of the factual world and the problem is that, within the framework of games, there is no unique way to do so; many different departures could provide an access to circumstances in which a deviation takes place. In some counterfactual worlds, deviations do not cause further deviations whereas they do in others. To find out the deviation-world closest to the actual world, the players need to agree upon a theory to analyze the relevant counterfactuals. The assumption that players share this theory implies that they also share a way of interpreting deviations. We will come to this point later on.

Finally, it is worth noticing that when S(m)=T(m) for every message off the equilibrium path, then there is no type that can be eliminated from the support of the beliefs of the second player. In this case the Intuitive Criterion results equivalent to the notion of sequentiality; that is, the criterion fails to refine the set of beliefs at nodes off-the-equilibrium path.

**2.3 Divine equilibria**

Consider the following game taken from Banks and Sobel [1]. Player 1 can be one of two types, called $t_1$ and $t_2$ with probabilities of 1/2 each. Each type has the same set of available messages; namely $m_1$ and $m_2$. The receiver has the same available actions after any of the messages: $a_1$ and $a_2$. The corresponding payoffs are depicted in Figure 3:

| $m_1$ | $a_1$ | $a_2$ | $m_2$ | $a_1$ | $a_2$ |
|---|---|---|---|---|---|
| $t_1$ | -3,3 | -6,0 | $t_1$ | -5,5 | -6,0 |
| $t_2$ | -3,3 | -11,5 | $t_2$ | -5,5 | -11,5 |

Figure 3

This game has two equilibrium outcomes:

$\{p(m_1/t_1) = p(m_1/t_2)=1;\ r\ (a_1/m_1)=r\ (a_1/m_2)=1\}$ and $\{p(m_2/t_1) = p(m_2/t_2)=1;\ r\ (a_1/m_2)=1\ \ r\ (a_1/m_1)=0\}$. Both equilibrium outcomes survive the Intuitive Criterion. However, Banks and Sobel claim that the equilibrium outcome that has both types sending message 2 is not sensible. The reason is that in order to support this equilibrium player 2 should believe that $t_2$ is more likely than $t_1$ and one can observe that whenever $t_1$ benefits from a response by player 2, $t_2$ benefits as well and not vice versa. We can describe this situation in the following manner: when $m_1$ is sent, the set of behavioral strategies by player 2 that outweighs the equilibrium payoff to $t_1$ contains the set of behavioral strategies by player 2 that outweighs the equilibrium payoff to $t_2$. Banks and Sobel assert that a sensible restriction to player 2's beliefs is that the probability of $t_1$ relative to that one of $t_2$ should increase when $m_1$ is received.

Their test can be summarized in the following manner. Let $A_G$ be the subset of P(A(m')) such that type $t\in T$ obtains a payoff at least as good as the equilibrium payoff denoted by u*(t) when he sends message m'. Formally,

$$A_G(m') = \{\alpha\in P(A(m')): u(t,m',\alpha) \geq u^*(t) \text{ for some } t\in T\ \}$$

This is the set of actions by player 2 that type t prefers to equilibrium actions if he sends message m'. Banks and Sobel assume that the receiver should believe that the sender does not expect to lose from a deviation. Therefore the receiver should believe that type t expects him to take an action from $A_G$.

For all actions in P(A(m')) let $\lambda$ be defined as:

$$\lambda(t,\alpha)=\begin{cases}1 & \text{if } u(t,m',\alpha) > u^*(t)\\ [0,1] & \text{if } u(t,m',\alpha) = u^*(t)\\ 0 & \text{if } u(t,m',\alpha) < u^*(t)\end{cases}$$

$\lambda(t,\alpha)$ represents the probability that $t\in T$ would send m' if he believed that m' would induce action $\alpha$, assuming that he could have obtained the equilibrium payoff u*(t). Now let $\Gamma(m',\alpha)$ be the set of player 2's beliefs over the set of types T consistent with player 2 taking action $\alpha$ in response to m' and type t obtaining u*(t) otherwise. Formally:

$\Gamma(m',\alpha) = \{\tau \in P(T): \exists\ \lambda(t)\in\lambda(t,\alpha)$ and $c>0$ such that $\tau(t)=c\lambda(t)\pi(t)\ \forall t\in T\}$.

Notice that this set is non empty if and only if $\alpha\in A_G$. Finally let

$\Gamma(A,m') = co[\cup_{\alpha\in A}\Gamma(m',\alpha)]$.

This set is empty only when $A_G \cap A$ is empty; this happens when there is no type who can strictly benefit from a deviation considering all possible responses by player 2. When $A_g$ is empty *any* conjecture supports the equilibrium. Otherwise, Banks and Sobel assert that it is not plausible for player 2 to hold beliefs outside $\Gamma(A,m')$ given the signal m'. Conjectures in $\Gamma(A,m')$ assign zero probability to types who can never benefit from a deviation with respect to their equilibrium payoffs. Moreover, when $A_G(t,m') \subset A_G(t',m')$ for $t,t'\in T$ then for all beliefs in $\Gamma(A,m')$ the ratio of the probability of t' given m' to the probability of t given m', is at least as great as $\pi(t')/\pi(t)$.

Finally we can present the iterative procedure introduced by Banks and Sobel:

Let $\Gamma_0 = P(T)$, $A_0 = P(A)$ and for $n>0$

$\Gamma_n := \Gamma(A_{n-1})$ if $\Gamma(A_{n-1})\neq\varnothing$ and $\Gamma_n := \Gamma_{n-1}$ otherwise.

$A_n := BR(\Gamma_n,m)$, $\Gamma^*=\cap_n \Gamma_n$, and $A^*=\cap_n A_n$.

A sequential equilibrium in a signaling game is *divine* if it is supported by beliefs in $\Gamma^*$. Returning to the game in Figure 2, the equilibrium outcome in which both types send $m_2$ is not divine[3]; in this case $\Gamma^*=\{\tau\in P(T): \tau(t_1)=\pi(t_1)=1/2\}$ and, as we already explained, these beliefs do not support player 2's off the equilibrium response in the case in which message $m_1$ sent.

---

[3] The only equilibrium outcome which is *divine* is the pooling equilibrium in which both types send message 1. In this case $A^*= a_1$.

This example also shows that the outcome of this procedure depends on the prior distribution $\pi$ over the set of types T. With the purpose of overcoming this limitation, Banks and Sobel redefine the set of beliefs that support the optimal response by player 2. Let $\Gamma^{**}$ be the intersection of every $\Gamma^{*}$ taken over all non degenerate priors on Sender types. A sequential equilibrium is *universally divine* if it is supported by beliefs in $\Gamma^{**}$. Naturally this is more restrictive than Divinity. In the game depicted in Figure 2, the pooling equilibrium in which message 2 is sent is divine only if $\pi(t_1)\leq 2/5$; however, this equilibrium is not universally divine, given that player 2 ought to believe that regardless of the prior the unexpected message comes from $t_1$.

### 2.4 The Intuitive Criterion and Divinity compared

Let us summarize the most important issues that regarding the methods presented in sections 2 and 3 have been discussed previously:

1) The interpretation of deviations as signals, which is an assumption in both the Intuitive Criterion and Divinity, builds upon the idea that deviations could emerge as a consequence of a rational decision.[4] As it was asserted in the previous section, this is one of the possible ways in which deviations can be interpreted. However, it is not clear whether this is the right way to think about deviations because it forces a causality that does need to hold: a deviation does not have to be rational because it was chosen by a rational player.

2) In the beer-quiche game, had a strong player 1 deviated from the quiche equilibrium he would have been better off in any of these two possible scenarios: *either* player 2 does not duel after beer *or* player 2 duels after quiche. The first case involves a deviation *off* the equilibrium path whereas the second a deviation *along* the equilibrium path and therefore a rejection to the assumption that players are playing the equilibrium under analysis. On the other hand, a deviation by the weak player 1 could be profitable *only if* he expects player 2 to deviate from his equilibrium strategy after *every* possible message.

This means that the weak type can be justifiably eliminated only when we assume that player 2 responds to equilibrium play with equilibrium strategies. In other words, we need to fix the equilibrium under consideration and proceed *as if the reasoning mechanism*, which is common knowledge among the players, *had no further consequences upon the decision to play the equilibrium strategies*. Both the Intuitive Criterion and Divinity build upon this

[4] In [10] we conjectured that Bennett's theory of counterfactuals [2] could be used to support the interpretation of deviations as acts of rational agents who may still be rational in scenarios where they deviate. This, we claim could support the hypothesis that deviations could be intentional. See section 2.4 for a brief comment about this way of thinking about deviations.

assumption. Moreover, once a player is eliminated, we consider player 2's best reply given the beliefs modified by this elimination. As it was already asserted, *if* under the quiche equilibrium beer signals a strong type *and* this implies that quiche signals a weak type *then* player 2 is better off by responding with a duel along the equilibrium path. This type of iteration is not possible under the methods described in sections 2 and 3 because they change player 2's response under equilibrium.

3) There are examples in which *every* type may potentially benefit from a deviation, even when the equilibrium is fixed. In this case the *Intuitive Criterion* yields no further refinement upon the set of sequential equilibria. In some of these cases, *Divinity* is capable of further restricting the set of sequential equilibria. To illustrate this consider the following example taken from Van Damme [12] :

| $m_1$ | $a_1$ | $a_2$ | $m_2$ | $a_1$ | $a_2$ |
|---|---|---|---|---|---|
| $t_1$ | 2,2 | 2,2 | $t_1$ | -1,5 | 3,0 |
| $t_2$ | 2,2 | 2,2 | $t_2$ | 0,0 | 4,1 |

Figure 4

Banks and Sobel's *Divinity* renders the equilibrium $\{p(m_2)/t_1)=p(m_2/t_2)=1; r(a_2/m_2)=1\}$ as the only divine equilibrium provided that $\pi(t_2) >0.5$. The other equilibrium outcome is $\{p(m_1)/t_1)=p(m_1/t_2)=1; r(a_1/m_1)=1\}$which is not divine. In order to illustrate this result, fix the pooling equilibrium in which player 1 plays $m_1$. Banks and Sobel's iterative procedure starts by updating beliefs in the following way: consider the set of possible responses by player 2 which would make each type better off. In this case $t_2$ gains by deviating to $m_2$ only if $r(a_2/m_2)\geq 0.5$. On the other hand, $t_1$ benefits from a deviation to $m_2$ only if $r(a_2/m_2)\geq 0.25$. This implies that $t_2$ gains by deviating whenever $t_1$ does. Divinity requires that the belief that $m_1$ comes from $t_2$ be at least equal to the prior probability of this type. However this would make player 2 deviate to $a_2$ which eliminates this pooling equilibrium as a candidate for Divinity.

Both equilibrium outcomes of this game pass the test or speech proposed by Cho and Kreps' whereas only $\{p_1(m_2)= p_2 (m_2)=1; r_{m2} (a_2)=1\}$ is divine.

Finally, it is worth noticing that *Divinity* depends upon the prior distribution over types. In the game depicted in figure 4 there is only one divine equilibrium provided that $\pi (t_2) >0.5$. However, when $\pi(t_2) \leq 0.5$ both equilibrium outcomes are divine.

## 3 Priors as off-the-equilibrium beliefs

Equilibria in signaling games of the sort described in section 1 can be typically of two different types: either pooling or separating. In the first case each type of player 1 chooses the same message and this implies that player 2 does not learn anything about the type of his opponent along the equilibrium path. This is due to the fact that, under this circumstance, Bayes' rule renders the conditional posterior of the types equal to the prior distribution. In the second case, messages perfectly signal player 1's type and in this way player 2 knows with certainty which type of player 1 he is facing.

Consider again the beer-quiche game. In order to support the equilibrium in which both types have beer for breakfast it is necessary that, if player 1 has quiche for breakfast player 2 duels with a probability of at least 0.5. In order to do so, player 2 needs to have beliefs that assign a probability of at least 0.5 to the weak type conditioned on the observation of quiche. In order to avoid the duel *along* the equilibrium path, player 2 needs beliefs that assign a probability of at least 0.5 to the event that he faces a strong player 1 conditioned on the observation of beer. Given that under this equilibrium $\tau_b^p(s):=\pi(s) = 0.9$, player 2 has no incentives to deviate along the equilibrium path. However, if player 2 uses the prior distribution over types as his beliefs in case of a deviation, he should not duel either and this clearly breaks the equilibrium under consideration.

Using this criterion and given the structure of this game, existence of Nash equilibrium would be guaranteed only if $\pi(s)=\pi(w)=0.5$. That is, existence is guaranteed only when the prior equals the threshold probability that makes player 2 indifferent between the two different responses, along and off the equilibrium path.

The equilibrium in which both types have beer for breakfast is supported only when quiche signals that a weak type is more likely than a strong type. When player 2 does not update his beliefs in this way, the weak type will have clear incentives to deviate to a breakfast of quiche. The pooling equilibria of this game requires that the two sets of player 2's beliefs that respectively support his response on and off the equilibrium path have only one element in common, which we called the threshold. In this situation and under the assumption that on and off-the-equilibrium path beliefs are equal to the prior distribution over the types, Nash equilibrium exists only when this prior distribution equals the threshold beliefs over types that makes player 2 indifferent between his available actions.

Consider the pooling equilibrium outcome in the game depicted in Figure 3 in which both types send message 1 and player 2 replies with $a_1$. Regardless of player 2's beliefs, $a_1$ is always a best response (at least as good as $a_2$) if he faces message 2. Therefore in this case any prior used as beliefs off-the-equilibrium path will support the equilibrium. Consider now the pooling equilibrium in which both types send $m_2$. If player 2 responds with $a_2$ off-

the-equilibrium path, then no type will wish to deviate. This implies that player 2 should have beliefs that attach a probability of at least 3/5 to type $t_2$. If player 2 uses the prior distribution as his beliefs, any prior distribution such that $\pi(t_2)\geq 3/5$ supports the equilibrium outcome under consideration.

### 3.1 The role of payoffs and priors in the existence of equilibrium

The drawback of the methodology outlined in the previous subsection is that existence of equilibrium can not be guaranteed. The restriction imposed upon the set of beliefs over the types in case of a deviation may be too strong, given the payoff structure of the game to allow for an equilibrium. When the prior distribution over the types is taken as the beliefs of the second player in a signaling game, existence of a pooling equilibrium is only guaranteed for a strict subset of priors if we fix the payoff structure. However, given that players decide upon their play by considering their expected utility it should also be noted that the payoff structure, albeit typically fixed, also plays a role. To analyze this interaction let us consider the basic case of two types of player 1, two messages and two replies by player 2:

| $m_1$ | $a_1$ | $a_2$ | | $m_2$ | $a_1$ | $a_2$ |
|---|---|---|---|---|---|---|
| $t_1$ | $u^1_{11},v^1_{11}$ | $u^1_{12},v^1_{12}$ | $t_1$ | $u^2_{11},v^2_{11}$ | $u^2_{12},v^2_{12}$ | |
| $t_2$ | $u^1_{21},v^1_{21}$ | $u^1_{22},v^1_{22}$ | $t_2$ | $u^2_{21},v^2_{21}$ | $u^2_{22},v^2_{22}$ | |

Figure 5

Without loss of generality assume that the pooling Nash equilibrium of this game is: $E=\{m_1, m_1, a_2, a_1\}$. Player 2 prefers $a_2$ to $a_1$ after receiving message $m_1$ if and only if:

$$v^1_{11}\,\tau_{m1}(t_1) + v^1_{21}\,[1-\tau_{m1}(t_1)] \leq v^1_{12}\,\tau_{m1}(t_1) + v^1_{22}\,[1-\tau_{m1}(t_1)] \qquad (3.1.1)$$

Along the equilibrium path $\tau_{m1}(t_1)=\pi(t_1)$. Therefore we can rewrite (3.1.1) as:

$$v^1_{11}\,\pi(t_1) + v^1_{21}\,[1-\pi(t_1)] \leq v^1_{12}\,\pi(t_1) + v^1_{22}\,[1-\pi(t_1)] \qquad (3.1.2)$$

On the other hand, player 2 prefers $a_1$ to $a_2$ after receiving message $m_2$ if and only if:

$$v^2_{11}\,\tau_{m2}(t_1) + v^2_{21}\,[1-\tau_{m2}(t_1)] \leq v^2_{12}\,\tau_{m2}(t_1) + v^2_{22}\,[1-\tau_{m2}(t_1)] \qquad (3.1.3)$$

Assuming that $\tau_{m2}(t_1)=\pi(t_1)$ we rewrite (3.1.3) as:

$$v^2_{11}\,\pi(t_1) + v^2_{21}\,[1-\pi(t_1)] \leq v^2_{12}\,\pi(t_1) + v^2_{22}\,[1-\pi(t_1)] \qquad (3.1.4)$$

Equations (3.1.2) and (3.1.4) provide two constraints for the values or $\pi(t_1)$ such that an equilibrium exists. It is clear that equilibrium exists, if only if the actual value of $\pi(t_1)$, satisfies both equations and this at least requires that the intersection of these two bounds be non empty.

Although the payoffs in equations (3.1.2) and (3.1.4) do not overlap, the assumption that posterior beliefs equal prior probabilities introduces a link between the responses to different messages received by the second player. In other words, although replies to every message are based upon different payoffs, they should be motivated by compatible beliefs.

In the beer quiche game (Fig 2) equations (3.1.2) and (3.1.4) provide the following bounds for the equilibrium in which both types drink beer: $\pi(s) \leq 0.5$ and $\pi(w) \geq 0.5$ respectively. Therefore equilibrium would exist if and only if $\pi(t_1)=0.5$. In the game depicted in Figure 3, if one considers the pooling equilibrium in which both types send $m_2$, equations (3.1.2) and (3.1.4) respectively require that $\pi(t_1) \geq 0$ and $\pi(t_1) \leq 0.4$. Thus equilibrium exists for $0 \leq \pi(t_1) \leq 0.4$.

**4 A variation of Banks and Sobel's *Divinity***

Consider the game in Figure 6 with $\pi(t_1)=0.9$ $\pi(t_2)=0.1$.

| $m_1$ | $a_1$ | $a_2$ | $m_2$ | $a_1$ | $a_2$ |
|---|---|---|---|---|---|
| $t_1$ | 0,0 | 0,0 | $t_1$ | -1,0 | 1,1 |
| $t_2$ | 0,0 | 0,0 | $t_2$ | -1,1 | 1,0 |

Figure 6

This game has two equilibrium outcomes: $\{p_1(m_1)= p_2(m_1) =1;\ r_{m2}(a_1)=1\}$ and $\{p_1(m_2)= p_2(m_2) =1;\ r_{m2}(a_2)=1\}$. Consider now the pooling equilibrium where both types send message $m_1$. No type can be eliminated by the *Intuitive Criterion*. Neither does *Divinity* refine the set of sequential equilibria because all of them are divine. The reason is that both types can potentially benefit from a deviation in exactly the same circumstances; that is, $A_G(t_1,m')=A_G(t_2,m')$. Therefore $\Gamma^*=P(T)$.

Consider now the following variation to the iterative procedure that defines divine equilibria:

$\underline{\lambda}(t,\alpha)= 1$ if $u(t,m',\alpha) \geq u^*(t)$ and

$\underline{\lambda}(t,\alpha)= 0$ otherwise.

$\underline{\Gamma}(m',\alpha) = \{\tau \in P(T);$ such that $\tau(t)=c\lambda(t)\pi(t)\ \forall t \in T$ and $c>0\}$.

Moreover when $A_G(t,m')$ is empty for all t in T and therefore $\underline{\Gamma}(A,m')=\varnothing$ assume that

$\underline{\Gamma}(m',\alpha) = \{\tau \in P(T);$ such that $\tau(t)=\pi(t)\ \forall t \in T\}$.

Calculate the remaining by following the iterative procedure outlined in Section 2.3 replacing $\Gamma$ and $\lambda$ by $\underline{\Gamma}$ and $\underline{\lambda}$ respectively.

The pooling equilibrium of the game depicted in Figure 6 in which both types play $m_1$ is supported by this variation of the procedure provided that $\pi(t_1)\leq 0.5$. With the modification presented above, player 2 uses the priors as his beliefs when he observes the off-the-equilibrium message $m_2$ and therefore decides to play $a_1$. This response deters both types from deviating.

### 4.1 The refinement to Divinity and Divinity compared

In this section, we compare the variation to Divinity just outlined with Banks and Sobel's iterative procedure. Let us consider again a signaling game in which there are two types of player 1, two available messages for each type and two available responses by player 2. Without loss of generality consider the following pooling equilibrium: $E = \{m_1, m_1, a_2, a_1\}$. In this equilibrium both types send $m_1$; player 2 responds with $a_2$ to this message and with $a_1$ to $m_2$.

Given the payoffs to the first player, four possible scenarios are feasible regarding the set of actions by player 2 that each type prefers to equilibrium actions:

i) $A_G(t_1)=A_G(t_2)=\varnothing$.

In this case no type can potentially gain by deviating relative to his equilibrium payoff and this implies that $\Gamma(A_{n-1})=\varnothing$ for all $n>0$. Moreover $\Gamma_n=P(T)$ for all $n>0$ and $\Gamma^*=\cap_n\Gamma_n = P(T)$. This means that every sequential equilibrium is Divine.

Regarding the variation presented in the previous subsection, $\underline{\Gamma}(m',\alpha)=\{\tau(t_1)=\pi_1;\ \tau(t_2)=\pi_2\}$. This implies that $\underline{\Gamma}^*$ is a singleton consisting of the prior distribution over the types. An equilibrium satisfies the test presented in section 4 if and only if the prior distribution over types satisfies the boundaries defined in (3.1.2) and (3.1.4).

ii) $A_G(t_1)=\varnothing$; $A_G(t_2)\neq\varnothing$ (alternatively $A_G(t_2)=\varnothing$; $A_G(t_1)\neq\varnothing$).

In this case, there is only one type who may benefit from a deviation and this leads player 2 to believe that the deviation certainly comes from this type: $\Gamma^*=\{\tau(t_1)=0;\ \tau(t_2)=1\}$(alternatively $\Gamma^*=\{\tau(t_1)=1;\ \tau(t_2)=0\}$). In this particular case, the refinement that Divinity imposes over sequentiality is equivalent to that imposed by the Intuitive Criterion. The beer-quiche game depicted in Figure 2 illustrates this case.

The variation of Divinity outlined in the previous subsection is equivalent in this case to Divinity; that is, $\underline{\Gamma}^*=\Gamma^*$.

iii)$A_G(t_1)=A_G(t_2)\neq\varnothing$.

In this circumstance the set of actions by player 2 that both types prefer to equilibrium actions coincides. This means that both types could potentially benefit from a deviation in the exact same situations. As in case i) $\Gamma^*=\cap_n\Gamma_n = P(T)$ and therefore Divinity is equivalent to sequentiality: an equilibrium is Divine if and only if it is sequential. In the game introduced in Figure 6 the set of equilibrium outcomes such that both types send $m_1$ is an example of this case.

The modified set of beliefs $\underline{\Gamma}^*$ by the second player is a singleton in this situation consisting of the prior distribution over the types as in case i). As before, an equilibrium satisfies the test presented in section 4 if and only if the prior distribution over types satisfies the boundaries defined in (3.1.2) and (3.1.4).

iv)$A_G(t_1)\neq A_G(t_2)$; $A_G(t_1)\neq\varnothing$; $A_G(t_2)\neq\varnothing$.

Although both types may potentially gain relative to their equilibrium payoffs, there are responses by player 2 that would induce a deviation by only one of the types. That is, either $A_G(t_1)\subset A_G(t_2)$ or $A_G(t_2)\subset A_G(t_1)$. The game depicted in Figure 3 illustrates the case in which $A_G(t_2)\subset A_G(t_1)$. As we already saw not every sequential equilibrium is Divine in this circumstance although every sequential equilibrium satisfies the Intuitive Criterion.

The variation of Divinity outlined in section 4 is equivalent in this case to Divinity; that is, $\underline{\Gamma}^*=\Gamma^*$.

**5 Concluding remarks**

There is no unique way to determine off-the-equilibrium path behavior and certain approaches are not always free of inconsistencies (see footnote 1). There are different ways through which the counterfactual world of a deviation could be reached, and every alternative involves the relaxation of at least one of the assumptions that hold in the equilibrium world, where players do not deviate.

There are on the one hand, the assumptions concerning the rationality of the players (including the possibility of their making a mistake either in calculation or implementation) and on the other, the assumptions concerning the amount of knowledge that players possess about the structure of the game and the rationality of their opponents. The revision or updating of beliefs, in the face of a deviation, will crucially depend upon the assumption being dropped to conceive that world and this, in turn, will determine whether the equilibrium under analysis is consistent with common knowledge of rationality.

A possible solution is to link the interpretation of deviations to the structure of the game when rationality is the last assumption that the theorist (and the players) may want to relax (this approach has the drawback of being an *ad hoc* procedure). Players may look first for "rational" reasons or signals behind a deviation. In the absence of such possibilities, that

is, when deviations could not conceivably lead to a potential gain, intentionality can not be compatible with rationality. In this case, deviations may be considered meaningless and it seems reasonable that players base their responses upon the information which is common knowledge in the game and proceed as if no further deviations were expected.

In [10], we conjectured that Lewis' theory of counterfactuals supports the hypothesis that deviations should be taken as thought experiments, in the sense of being options available to the players that they can scrutinize even though, under equilibrium, they constitute irrational choices. To conceive a situation, or world, in which a deviation might have occurred, and according to Lewis' theory, we could think of a tremble or mistake, uncorrelated with further play[5]. That is, in the counterfactual scenario of a deviation, the player *would have acted irrationally* –a requisite to conceive this possibility- however, this irrationality would not be interpreted as an *intrinsic disposition* of the player, which could influence his further play in other worlds or states. This irrationality occurs in this scenario and is not necessarily correlated with the thought experiment needed to conceive *other* worlds where other deviations by this player occur. Of course, another possibility is to deduce that the player is more likely to behave irrationally again, in the sense or his reasoning capabilities, and that this pattern that may bring correlations between scenarios where this deviator plays again. This brings us to the next approach.

Following Harsanyi [6] and along the lines of Kreps *et al*. [8], one could also assume that at least one piece of information concerning the structure of the game is not common knowledge. For instance, it can be supposed that players (or some "types") are guided by different payoffs from the ones their opponents had expected them to have[6]. Irrational players do not actually behave irrationally within the theory of the game; they just have payoffs that make them act as if they were. Of course this is just a technical device aimed at incorporating the possibility of irrational play in the actual world, where the players conform to the equilibrium. This approach however, would add more types to our signaling game. It is worth noticing that, in order to be effective and complete, this model should include all relevant types to explain all possible sources of deviations, including mistakes in implementation. Otherwise, there will be a need for an "exogenous" belief revision policy.

---

[5] The possibility of making a mistake while implementing the choice must always have some positive probability. People do make mistakes unintentionally. Of course, the thesis that they make these mistakes repeatedly or systematically can not be defended. After many (how many?) deviations, a player may better be seen as irrational in a mode fundamental sense than in mere implementation. This will depend on the particular game and context and could be settled experimentally.

[6] This is the approach supported by Kreps [9] (pp. 489-496) and Skyrms & Woodruff [9].

Let us consider again the case, where some player may potentially benefit from not conforming to his equilibrium strategy -on the assumption that his deviation, by signaling his intention, will trigger further deviations. In this situation, deviations would signal either the deviator's future play, or reveal some information, which was not common knowledge before. The deviation would have a logic the opponent is supposed to deduce.

Signaling games constitute a good example to apply this criterion. We have proposed that signals off-the-equilibrium path exist only, or at most, in the type of situations described in the previous paragraph. The refinements considered in this paper, namely The *Intuitive Criterion* and *Divinity*, refine the set of sequential equilibria in some of these circumstances. The first refinement is effective when there are players whose deviations could be rationally explained. The second refinement is effective not only in this circumstance, but also when the players who may deviate, can be separated in terms of the responses that they would prefer after a deviation. In the cases in which this is not possible, we have proposed a variation of this second refinement aimed at determining the beliefs of the uninformed player. A case has been made in the present paper, for the use of prior probabilities as the ex-post beliefs after a deviation. This could be justified by the fact that the prior distribution over types is a piece of information which is common knowledge ex-ante and guides players' responses along the equilibrium path. Nevertheless, we do not claim that this is, or should be, *the only way* to determine beliefs in the face of a deviation. It is just one possibility that minimizes the differences between the equilibrium-world and the deviation-world, and it could refine the set of equilibria, depending on the structure of the game. Our interest was to explore, in terms of the literature of equilibrium refinements, the consequences of such belief revision policy.

**References**


[1] Banks, J. & Sobel, J. "Equilibrium Selection in Signaling Games" Discussion Paper # 85-9, mimeo, Department of Economics University of California, San Diego.

[2] Bennett, J. "Counterfactuals and Temporal Direction", The Philosophical Review 93 (1984), pp. 57-91.

[3] Cho, I.K. & Kreps, D. "Signaling Games and Stable Equilibria" The Quarterly Journal of Economics, vol.CII; May 1987 Issue 2.

[4] Fudenberg, D. and Tirole, J. Game Theory MIT Press, 1991.

[5] Harper, W. "A sketch of some recent developments in the theory of conditionals" in IFS edited by Harper, W., Stalnaker,R. and Pearce, G. D.Reidel Publishing Company.

[6] Harsanyi, J. "Games with incomplete information played by Bayesian Players" Parts (I, II and III) Management Science 14 159-182, 320-334, 486-502 (1967).

[7] Jackson, F. "A Causal Theory of Counterfactuals", Australian Journal of Philosophy 55 (1977).

[8] Kreps, D. Milgron P. Roberts J. and Wilson, R. "Rational cooperation in the finitely repeated prisoners' dilemma" Journal of Economic Theory 27: 245-252, (1982).

[9] Kreps, D. "A Course in Microeconomic Theory" Princeton University Press 1990.

[10] Rodríguez Mariné, G. "The Backward Induction Solution to the Centepede Game" PhD thesis, University of California, Los Angeles, Department of Economics, 1995.

[11] Skyrms, B. & Woodruff, P. "Bayesian Conditionals for Bayesian Game Theory" for the conference on "Counterfactual Thought Experiments in World Politics: Logical, Methodological and Psychological Perspectives" Berkeley, CA Jan. 13-14, 1995.

[12] Van Damme, E. "Stable equilibria and forward induction" Journal of Economic Theory 48, 476-496.

[13] Van Damme, E. "Stability and Perfection of Nash Equilibria" Springer-Verlag 1987,1991.