Forward Induction as Confusion over the Equilibrium Being Played Out

26 January 1991/11 September 2007

Eric Rasmusen

Abstract

The Nash equilibrium of a game depends on it being common knowledge among the players which particular Nash equilibrium is being played out. This common knowledge arises as the result of some unspecified background process. If there are multiple equilibria, it is important that all players agree upon which one is being played out. This paper models a situation where there is noise in the background process, so that players sometimes are unknowingly at odds in their opinions on which equilibrium is being played out. Incorporating this possibility can reduce the number of equilibria in a way similar but not identical to forward induction and the intuitive criterion.

Eric Rasmusen, Dan R. and Catherine M. Dalton Professor, Department of Business Economics and Public Policy, Kelley School of Business, Indiana University. Visitor (07/08), Nuffield College, Oxford University. Office: 011-44-1865 554-163 or (01865) 554-163. Nuffield College, Room C3, New Road, Oxford, England, OX1 1NF. Erasmuse@indiana.edu. http://www.rasmusen.org. Copies of this paper can be found at: http://www.rasmusen.org/papers/rasmusen-subgame.pdf.

I would like to thank Emmanuel Petrakis and participants in the University of Chicago Theory Workshop and the University of Toronto for helpful comments.

1. Introduction

The idea here will be: if players observe actions by player Smith that are compatible with Nash equilibrium E_1 but not with Nash equilibrium E_2, they should believe that Smith will continue to play according to equilibrium E_1, even if they themselves were earlier intending to play according to equilibrium E_2.

Background: Equilibrium in the Expensive-Talk Game

Let us use the Expensive-Talk Game (also known as the Money-Burning Game) as an example for discussing equilibrium concepts.

Expensive-Talk Game I

1. The Man chooses Talk and say "The strategies chosen will be (*Fight, Fight*)" at cost $c < 2$, or choose Silence at cost 0, observed by the Woman.

2. The Man and Woman simultaneously choose *Fight* or *Ballet*, which add the amounts in Table 1 to their payoffs.

		Woman	
		Fight	Ballet
Man	Fight	**3,1**	0,0
	Ballet	0,0	**1,3**

Payoffs to: (Man, Woman).
(Nash equilibrium payoffs are in boldface.)
Table 1: The Battle of the Sexes

Move (2) is the game known as the Battle of the Sexes, a coordination game in which the two players wish to choose the same action but have different preferences over which action to choose. The players are a man who wishes to go to a prizefight and a woman who wishes to go to a ballet, both of whom would also like to attend an event together. The game

has two pure-strategy equilibria, (F:*Fight, Fight*) and (B:*Ballet*, *Ballet*), and a mixed-strategy equilibrium M, in which each player picks his preferred action with probability .75.

Even in the Battle of the Sexes by itself, there are several arguments that can narrow down the number of equilibria.

1. Exclude mixed-strategy equilibria, as being more complicated and special. This excludes equilibrium M.

2. Exclude pareto-dominated equilibria, since the players would endeavor to avoid them. This excludes equilibrium M, which has an expected payoff of xxx.

3. Exclude asymmetric equilibria, in which one player has a higher payoff than the other, since they would be harder to coordinate upon. This excludes equilibria F and B.

None of these principles are compelling, but for our illustration we will adopt principle (1) and exclude mixed-strategy equilibria, as is commonly done when pure-strategy equilibria exist. [1]

The Expensive-Talk Game precedes the Battles of the Sexes with a move in which the man can choose $Talk$ to try to convince the woman that he will pick $Fight$ in the Battle of the Sexes subgame. This has the following Nash equilibria (all of which are subgame perfect):

E1: (Man: S, B—S, B—T) (Woman: B—S, B—T). Outcome: SBB.

E2: (Man: S, F—S, F—T) (Woman: F—S, F—T). Outcome: SFF.

E3: (Man: S, F—S, B—T) (Woman: F—S, B—T). Outcome: SFF.

E4: (Man: T, B—S, F—T) (Woman: B—S, F—T). Outcome: TFF.

Seen as Bayesian games, though, in which players are required to

[1] If c took a small value, M could be used as a punishment to support peculiar Nash equilibria such as TBB in the Expensive-Talk Game. The equilibrium with that outcome is Man: TB, Woman:$(B|T, M|S)$. The man is willing to bear the cost of $Talk$ because if he deviated, he would punished with MM in the subgame, which is even worse than BB for him if for $c < 1.75$.

behave rationally in light of their prior beliefs as updated using Bayes's Rule, the statements of the equilibria are incomplete. In addition to the strategies listed, prior to the game starting the players implicitly hold the belief that with probability 100% the particular equilibrium is the one the other player will play out. Thus, the priors for E1 are:

Man: With probability 1, the Woman will respond to S by choosing B.

Woman: With probability 1, the Man will choose S, after which he will choose B.

But what is the woman to believe if she observes the man choosing Talk? That is an impossible event, in light of her prior, and Bayes's Rule provides no way to update a prior of 0 or 1. Thus, to fully specify the equilibrium, the modeller must specify the woman's posterior after she observes $Talk$. Here, an out-of-equilibrium belief that supports the equilibrium is

Woman's out-of-equilibrium belief: If the man chooses S, then with probability one he will choose B.

Not all beliefs support the equilibrium. An equally rational belief, but one that does not support the equilibrium, is

Woman's out-of-equilibrium belief: If the man chooses S, then with probability .5 he will choose B.

The concept of perfect bayesian equilibrium, in which players are required to behave rationally in light of their prior beliefs and in which out-of-equilibrium beliefs are specified by the modeller, is usually applied to games of incomplete information, in which the priors are about an initial move by Nature choosing the type of a player. It is equally applicable to games of complete information. We implicitly assume, however, that the out-of-equilibrium belief is whatever belief will support the specified subgame-perfect Nash equilibrium.

In some games, however, including the Expensive-Talk Game, people feel uncomfortable with the out-of-equilibrium beliefs.

FI equilibria. Only SFF is an equilibrium. (1) SBB is not an FI

outcome, because it is supported only by the equilibrium (SB,B), and in that equilibrium, TB is dominated. If we drop TB, then B is no longer a perfect strategy for the woman: if the man chooses TF, she must rationally respond with F, so SBB is not self-enforcing when dominated strategies are dropped. (2) TFF is not an FI outcome because SB is dominated in it. If we drop SB, however, then if the man picks S, that indicates to the woman that the man has chosen SF, so she will respond with F. Thus, TFF is not self-enforcing when dominated strategies are dropped. (3) SFF is an FI outcome because if we drop the man's TF, TB, or SB, or the woman's $(B|S, F|T)$ or B, that does not stop SFF from being self enforcing.

The content of the announcement is unimportant, only its cost. Even if the man spends 1.5 to say "The strategies to be played out are (*Ballet, Ballet*)" the woman will believe that he means to convey to her that the equilibrium is (*Fight, Fight*); It's not what you say, it's how you say it. But as we have just seen, silence can then convey the message as effectively as the announcement and more cheaply.[2] Given that the man can guarantee a favorable outcome by *Talk*, he would refrain from talking only if talking were unnecessary, which is true only if he thinks that the woman will pick $Fight$ anyway. If he thinks that, he will choose $Fight$ himself and the woman will also wish to choose $Fight$. So the man's silence also communicates that he will choose $Fight$. The key to the success of the "strong, silent type" is that he have the the *option* of sending a costly message; It's not what you say; it's whether you can say it.

This line of reasoning strikes some people as so unintuitive as to condemn the whole idea of forward induction, but the strangeness of the result may be due to its lack of robustness with respect to asymmetric information. If with some probability the rules of the game do not allow the man to make announcements, the woman can interpret silence as indicating that the man *cannot* talk, rather than that he feels no need to talk, in which case TFF remains an equilibrium. In the extended game, if Nature sends the message to both that the equilibrium is for the man to talk if he is able and to be silent otherwise and for the both players to choose $Ballet$ if the man is silent, then the man cannot induce the woman to choose $Fight$ by staying silent. A similar argument for the equilibrium status of TFF could

[2]The argument for the effectiveness of silence in this kind of game can be found in Ben-Porath & Dekel (unpublished) and Van Damme (1989).

be based on the woman not knowing the man's preference exactly, assigning some probability between 0.5 and 1 that the man prefers the prizefight. A TFF equilibrium would then exist in which the man chooses $Talk$ if and only if he prefers the prizefight. In the extended game, if the woman received the message for that TFF equilibrium, she would interpret the man's silence to mean that he preferred the prizefight; if he actually prefers the ballet, he must talk. Thus, with even a little asymmetric information it is hard to rule out TFF and forward induction only rules out SBB.[3]

2. Give my new definition.

1. Determine the pure-strategy Nash equilibria of the game. Here, their outcomes are E1: SFF, E2: SBB, and E3: TFF. In full, these include out-of-eq. beliefs too.

E1: (Man: S, F—S, F—T, OOB: T means both players will choose F) (Woman: F—S, F—T, OOB: T means both players will choose F).

2. Nature begins the game by sending each player a message of which equilibrium is to be played out: E10, E20, etc.

3. With probability (1-epsilon), all players hear the same message. With probability epsilon, the messages differ. We will not restrict the way they might differ, because we will look for equilibria robust to all possibilities. For example, we could assume that if E10 is the high-probability message, then

A. With probability epsilon, Players 1, 2, 3 hear E10 but Player 4 hears E20, E30, or E40 with 1/3 probability each.

or

B. With probability epsilon, Players 1 hears E10, Players 2 and 3 hear E20, and Player 4 hears E30.

4. We will only consider Nash equilibria of the metagame in which players play out the equilibrium they hear. (This is not the same as assuming that in the main game they play the equilibrium they hear: doing so must be a Nash equilibrium of the metagame too.) Such equilibria

[3]This argument can be found in footnote X of Van Damme (1989).

definitely exist, but other equilibria exist too, such as "Play E40 regardless of the message you hear."

5. If one of the set E10, E20, etc. now fails to be Nash equilibria in the metagame for some possible epsilon-specification, drop it.

6. Iterate using the surviving set E11, E21, etc.

Confusion-proof Equilibria in ET I. SFF is the only equilibrium. (1) Suppose SBB were a confusion-proofness equilibrium and the man deviated from it by picking T. The woman concludes that either she misheard (and nature chose TFF), or the man misheard (and nature chose SBB). In either case, the man will pick F, so the woman does also. The man's deviation was profitable. (2) Having disposed of SBB, suppose the equilibrium outcomes are TFF and SFF. If Nature sends the message TFF to each player, but the man remains silent, the woman deduces that he received SFF and she chooses F. The man can safely choose F himself, and the deviation has been profitable to the man, so TFF is broken as an equilibrium outcome.[4]

Expensive-Talk Game II: Incomplete Information

1. Nature chooses F or B for the man, unobsreved by woman.
2. The Man chooses Talk, at cost c, or Silence, at cost 0.
3. The Woman chooses *Fight* or Ballet.

We clearly need PBE for this.

Consider variant 3, the transition between the two, if these seem unreasonable. The order of moves is:

Expensive-Talk Game III: Post-Start Asymmetric Information

1. Man chooses F or B, unobsreved by woman.
2. The Man chooses Talk, at cost c, or Silence, at cost 0.
3. The Woman chooses *Fight* or *Ballet*.

In choosing a particular strategy combination to be a game's equilibrium, three criteria are generally accepted. First, the strategy

[4]I ruled out mixed-strategy equilibria, but that is not a necessary part of the concept. Suppose we allowed tehm. Then, there are five additional Nash equilibria, (T, M—S, M—S, B—T, B—T), (T, M—S, M—S, F—T, F—T), (S, F—S, F—S, M—T, M—T), (S, B—S, B—S, M—T, M—T), (S, M—S, M—S, M—T, M—T). CP will knock out the Silence equilibria. CP doesn't help more, even with iteration. FI probably does.

combination ought to be Nash— every player's strategy should be a best response to the other players' strategies. Second, it ought to be subgame perfect—for every subgame the relevant portions of the strategy combination should be Nash. Third, it should be a perfect Bayesian equilibrium—remaining Nash when the players follow Bayes Rule and some set of out-of-equilibrium beliefs assigned by the modeller.

The Nash property is the most fundamental of these three criteria, and it is attractive because of its consistency—every player's behavior and beliefs are consistent with every other player's. But there can be multiple Nash equilibria, and with multiple equilibria this consistency is a less compelling argument for Nash equilibrium. In equilibrium X a player may not be able to profit by deviating from action a_x, and in equilibrium Y he may not be able to profit by deviating from a_y, but how is he to know that the other players are playing X and not Y? Anyone who finds Nash equilibrium plausible must believe in some unmodelled background process by which the players come to know which particular equilibrium is being played out.

When there is just one Nash equilibrium, the background process can be simply that players realize that only one combination of strategies is consistent, but when there are many equilibria the background process is more mysterious. Games with multiple equilibria are common. Signalling games, in particular, are prone to multiple equilibria, because the concept of perfect Bayesian equilibrium allows many out-of-equilibrium beliefs, and a host of equilibrium refinements have been developed to reduce the number of equilibria (see Kohlberg [unpublished] or Van Damme [1987] for references). But multiple equilibria are a problem even in games of symmetric information, where the multiplicity does not arise from ignorance of the players' "types," but from ignorance of what actions they have taken in the course of the game. The most common examples are coordination games, in which the players wish to coordinate with each other by choosing the same actions. It would be desirable to find an equilibrium refinement that would have bite whether the information is asymmetric or not, and the presence of the problem in games of symmetric information suggests that it is not just a problem of out-of-equilibrium beliefs about types.

The great contribution of Harsanyi (1967) was to suggest that games

of incomplete information, in which a player is not sure whether he is playing in game X or game Y, could be remodelled as games of complete information in which Nature moves first and selects X or Y with known probabilities. Could the same be done for a game in which the players are uncertain over whether they are playing out equilibrium X or equilibrium Y? To do so would contradict a fundamental assumption of Nash equilibrium, that the equilibrium being played out is common knowledge and all the players' strategies are consistent, and if the probabilities of both X and Y are substantial this would not lead to sensible results. But we could hope that an equilibrium would be robust to having a little bit of uncertainty about whether it is X or Y that is being played, especially since we are somewhat hazy on how it becomes common knowledge which particular equilibrium is to be played out. This, after all, is the hope behind the "tremble" justification for perfect equilibrium, and it is easier to imagine players being confused over the difficult problem of deciding which equilibrium is being played than to imagine them blundering by choosing obviously unprofitable actions. As with trembles in actions, uncertainty over the equilibrium might reveal some Nash equilibria not to be robust, thus providing a way to refine the equilibrium concept. If the modeller believes that there is any uncertainty in the minds of players as to which equilibrium is being played out, his preferred equilibria should be robust to the presence of that uncertainty

Besides the Tremble and Asymmetric Information approaches the refinements there is a third approach, the Equilibrium Forcing approach. In it, a player chooses an action to indicate to the other players that he intends to play a particular strategy, in an effort to force them to change their expectations.

The present article takes the Asymmetric Information line of attack. To represent the background process of equilibrium selection, two initial moves by Nature will be added to the start of whatever game is under consideration: a first move in which an equilibrium X is selected and a second in which each player is sent a message announcing the equilibrium. With a small probability, the message is not X but some other equilibrium. The players must therefore take into account the possibility of confusion over which equilibrium is being played out. This provides an opportunity for their beliefs about the equilibrium of the original game to change and to

be manipulated by deviations.

The new refinement's implications are close but not identical to those of the existing idea of "forward induction," an Equilibrium Forcing idea. Elon Kohlberg & Mertens (1986) introduced forward induction as part of the refinement they call "stability," and it is used in the "intuitive criterion" which In-Koo Cho & David Kreps (1987) apply to signalling games. Eric Van Damme (1989) and a variety of unpublished studies—by Kohlberg, Okuno-Fujiwara & Andrew Postlewaite, Martin Osborne, and Ben-Porath & Dekel— have shown that forward induction has interesting implications even without the other criteria that compose "stability" and that forward induction can reduce the number of equilibria even in games of symmetric information. Forward induction has been commonly defined in terms of iterated deletion of dominated equilibrium strategies but commonly justified in terms of logical deductions made by the players. The intuition is that after player Smith takes an action that could benefit him if and only if player Jones held belief Y, Jones, realizing this, will adopt belief Y. Such an intuition violates the fundamental assumption of Nash equilibrium—that the information structure and the identity of the equilibrium to be played out are common knowledge among the players. Defining forward induction in terms of iterated dominance skirts around that issue; FI' will meet it head on.

The Problem to be Addressed: The Expensive-Talk Game and the Twice-Repeated Battle of the Sexes

To show the problem, I will lay out two variants of the Battle of the Sexes, a coordination game in which the two players wish to choose the same action but have different preferences over which action to choose. The players are a man who wishes to go to a prizefight and a woman who wishes to go to a ballet, both of whom would like to attend an event together. They make their choices simultaneously, and the payoffs are shown in Table 2. The game has two pure-strategy equilibria, (*Fight, Fight*) and (*Ballet, Ballet*).[5]

[5]There is also a mixed-strategy equilibrium, in which each player picks his preferred action with probability .75. If we denote the equilibrium mixed strategy for the Ballet-Fight subgame as M, then if c took a small value, M could be used as a punishment to support peculiar perfect Bayesian equilibrium such as TBB in the Expensive-Talk Game.

		Woman	
		Fight	Ballet
Man	Fight	**3,1**	0,0
	Ballet	0,0	**1,3**

Payoffs to: (man, woman).
(Nash equilibrium payoffs are in boldface.)

Table 2: The Battle of the Sexes.

		Woman			
		F	B	F|T, B|S	B|T, F|S
Man	Silence, Fight (SF)	**3,1**	0,0	0,0	**3,1**
	Silence, Ballet (SB)	0,0	**1,3**	1,3	0,0
	Talk, Fight (TF)	3-c,1	-c,0	**3-c,1**	-c,0
	Talk, Ballet (TB)	-c,0	1-c,3	-c,0	1-c,3

(Pure-strategy perfect equilibria when $c = 1.5$ are in boldface.)

Table 3: The Expensive-Talk Game

Variant 1: The Expensive-Talk Game. This variant precedes the simultaneous-move Battle of the Sexes with a single costly announcement by the man. If he chooses to talk, his announcement costs him $c = 1.5$ units of payoff (in contrast to the "cheap-talk" of Farrell [1987] in which $c = 0$).

The perfect equilibrium pure-strategy outcomes are SFF, SBB, and TFF.

The problem is with the SBB equilibrium. In it, the woman expects the man to be silent, because it would be irrational for him to uselessly

The equilibrium with that outcome is Man: TB, Woman:$(B|T, M|S)$. The man is willing to bear the cost of *Talk* because if he deviated, he would punished with MM in the subgame, which is even worse than BB for $c < 1.75$.

incur a cost, given that his message would not persuade her that he really mean to choose *Fight*. But what if he *does* choose *Talk*? What should the woman think? Within the model this is an impossible event. When something impossible within a person's frame of reference occurs in the real world, ordinarily one does not simply blink and go on as if no miracle had occurred. Rather, one rethinks one's frame of reference.

One reaction would be for the woman to question whether she really heard a message or was hallucinating. Another would be to question whether the man was rational. A third would be to suppose that the man had chosen *Talk* inadvertently, by a "tremble". All three of these would lend some support– but not confidence— to *Ballet* as a best response for the Woman.

A fourth reaction would be that the man was trying to indicate to her that he was trying to shift her expectation for his second-period action to *Fight* and that he expected to succeed in that attempt, so he was going to choose *Fight* himself. In that case, *Fight* is the woman's best response.

A fifth reaction would be that the man mistakenly thought that TFF was the equilibrium both players were supposed to be playing out. Thus, he thought that if he chose *Silence* then the woman would choose *Ballet*, but if he chose *Talk* she would choose *Fight*. If that is what he believes, the woman had better choose *Fight*.

The fourth idea— Equilibrium Forcing– is the idea behind Forward Induction as conventionally defined. The fifth idea– Incomplete Information about Equilibrium Expectations– is what I want to pursue in this paper.

Discussion. The content of the announcement is unimportant, only its cost. Even if the man spends 1.5 to say "The strategies to be played out are (*Ballet, Ballet*)" the woman will believe that he means to convey to her that the equilibrium is (*Fight, Fight*); It's not what you say, it's how you say it. But as we have just seen, silence can then convey the message as effectively as the announcement and more cheaply.[6] Given that the man can guarantee a favorable outcome by *Talk*, he would refrain from talking only if talking were unnecessary, which is true only if he thinks that the

[6]The argument for the effectiveness of silence in this kind of game can be found in Ben-Porath & Dekel (unpublished) and Van Damme (1989).

woman will pick *Fight* anyway. If he thinks that, he will choose *Fight* himself and the woman will also wish to choose *Fight*. So the man's silence also communicates that he will choose *Fight*. The key to the success of the "strong, silent type" is that he have the the *option* of sending a costly message; It's not what you say; it's whether you can say it.[7]

Variant 2: The Twice-Repeated Battle of the Sexes. Let the Battle of the Sexes be repeated twice, with no possibility of announcements. The players know the outcome of the first repetition when choosing their moves for the second.

The perfect equilibrium pure-strategy outcomes are BB-BB, BB-FF, FF- FF, and FF-BB.

Here, Equilibrium Forcing will eliminate BB-FF and FF-BB, while Incomplete Information about Equilibrium Expectations will not eliminate any of the equilibria.

The Expensive-Talk Game. The FI and FI' outcomes are SFF. This game will illustrate two steps of iteration of the extended game, and the idea that the absence of a message can be just as meaningful as a message.

FI equilibria. Only SFF is an equilibrium. (1) SBB is not an FI

[7] This line of reasoning strikes some people as so unintuitive as to condemn the whole idea of forward induction, but the strangeness of the result may be due to its lack of robustness with respect to asymmetric information. If with some probability the rules of the game do not allow the man to make announcements, the woman can interpret silence as indicating that the man *cannot* talk, rather than that he feels no need to talk, in which case TFF remains an equilibrium. In the extended game, if Nature sends the message to both that the equilibrium is for the man to talk if he is able and to be silent otherwise and for the both players to choose *Ballet* if the man is silent, then the man cannot induce the woman to choose *Fight* by staying silent. A similar argument for the equilibrium status of TFF could be based on the woman not knowing the man's preference exactly, assigning some probability between 0.5 and 1 that the man prefers the prizefight. A TFF equilibrium would then exist in which the man chooses *Talk* if and only if he prefers the prizefight. In the extended game, if the woman received the message for that TFF equilibrium, she would interpret the man's silence to mean that he preferred the prizefight; if he actually prefers the ballet, he must talk. Thus, with even a little asymmetric information it is hard to rule out TFF and forward induction only rules out SBB. This argument can be found in footnote X of Van Damme (1989).

outcome, because it is supported only by the equilibrium (SB,B), and in that equilibrium, TB is dominated. If we drop TB, then B is no longer a perfect strategy for the woman: if the man chooses TF, she must rationally respond with F, so SBB is not self-enforcing when dominated strategies are dropped. (2) TFF is not an FI outcome because SB is dominated in it. If we drop SB, however, then if the man picks S, that indicates to the woman that the man has chosen SF, so she will respond with F. Thus, TFF is not self-enforcing when dominated strategies are dropped. (3) SFF is an FI outcome because if we drop the man's TF, TB, or SB, or the woman's $(B|S, F|T)$ or B, that does not stop SFF from being self enforcing.

FI' equilibria. SFF is the only equilibrium. (1) Suppose SBB were an FI' equilibrium and the man deviated from it by picking T. The woman concludes that with probability .5 she misheard (and nature chose TFF), and with probability .5 the man misheard (and nature chose SBB). In either case, the man will pick F, so the woman does also. The man's deviation was profitable. (2) Having disposed of SBB, suppose the equilibrium outcomes are TFF and SFF. If Nature sends them message TFF to each player, but the man remains silent, the woman deduces that he received SFF and she chooses F. The man can safely choose F himself, and the deviation has been profitable to the man, so TFF is broken as an equilibrium outcome.

FI equilibria. Only BB-FF and FF-BB remain. Consider BB-BB, which gives the man a payoff of 1+1. The strategy for the man of playing Fight in the first round and Ballet in the second is dominated in this equilibrium; it could not profit him even if the woman also deviated in the second round. But when that strategy is eliminated, then if the man plays Fight the woman can conclude that he is playing the strategy of Fight in the first round and Fight in the second round. She will respond by switching to Fight in the second round, and the outcome will be FB-FF, which gives the man a payoff of 0+3. The man's deviation is profitable, so BB-BB cannot be an equilibrium outcome. Since the game is symmetric, FF-FF cannot be an equilibrium outcome either.[8]

BB-FF, on the other hand, is an FI equilibrium. Deviation by the man in the first round could never be profitable, because he obtains his desired outcome in the second round even without deviation. Deviation by

[8]I have taken this example from Van Damme (1989).

the woman in the first round is unprofitable because she would have to give up the BB payoff of that round.

FI' equilibria. BB-BB, BB-FF, FF-FF, and FF-BB are all FI' equilibria. Suppose that BB-BB were the equilibrium chosen by Nature in the extended game, and that both players received the message without garbling, but the man deviates to Fight in the first round. The woman's equilibrium interpretation of this is that the man heard either FF-FF or FF-BB from Nature, with equal probability. The expected value of her responding with Fight is therefore $0.5(1) + 0.5(0)$, and the expected value of responding with Ballet is $0.5(0) + 0.5(3)$. She responds with Ballet, and the man's deviation from BB-BB is unprofitable to him. Parallel reasoning shows that FF-FF is an FI' equilibrium.

Discussion. This example distinguishes two intuitions that are at work in forward induction. The FI' intuition is that players are uncertain over which equilibrium is being played out, and a deviation shows lack of synchronization. The FI intuition is that players might be trying to disrupt the normal play of the game. If BB-BB is the agreed equilibrium, the man might nonetheless deviate with Fight in the first round, an action which conveys the message, "I know we were supposed to play BB-BB, but I prefer FF, and I think I can make you choose Fight in the second round. To show my conviction, I have played Fight in the first round, an action which would be worse than useless if it did not convince you to switch to Fight. Maybe I am wrong and you will choose Ballet anyway, but I myself am choosing Fight again." It does not matter why the man has this belief that he can induce the woman to switch to Fight; whatever his reasoning, if he himself is convinced by it he will be choosing Fight in the second round, and that makes Fight the best response for the woman.

In FI' the player is taking strategic advantage of uncertainty in the background process that chooses the equilibrium to be played out, whereas in FI the player is taking strategic advantage of the background process by which players decide how to respond to what are known to be deviations from rationality.

3. Uncertainty over the Equilibrium to be Played Out

Let us first define a "metagame".

Metagame (F, C) is a game which precedes the original game with a move in which Nature uses a distribution function $F(C)$ to select one of the $N \geq 1$ strategy combinations C with positive probability and send a message to each player describing that single chosen strategy combination. With infinitesimal probability ε, however, a player is sent a message chosen from one of the other $(N-1)$ strategy combinations using distribution G. (we could make ε, and G different for each F if we wanted).

Nash assumption: An equilibrium is a strategy combination in some metagame in which no player has incentive to deviate from the strategies in his message, given that he expects the other players not to deviate.

A metagame could have some non-Nash strategy S as the only one in C, but S would not be an equilibrium, because some player would deviate from it. The G assumption plays no role in Nash equilibrium. (or does it– what about weak domination?)

FI assumption: If a strategy combination S is an equilibrium, then for any given metagame (F', C'), S could be included in C' and F' adjusted so that no player has incentive to deviate from the strategies in his message, given that he expects the other players not to deviate.

If we start with Silence, Prizefight, Prizefight as the only element of C, and add Talk, Prizefight, Prizefight, then under any F, the man will want to deviate to SPP. So TPP is not an equilibriumj.

If we start with C= (SPP, TPP) and add SBB, the man will want to deviate to TPP. So SBB is not an equilibrium.

If we start with C = (SBB, TPP) and add TFF, TFF is still viable. This is true for any metagame if we are allowed to choose F (or even not, in this case), so TFF is an eq.

Use the Van Damme example, with just two players in an expensie talk BS game, in this section.

Then, do the twic-repeated PD.

Global games should be discussed. It is a stronger concept, working even in a static game. There, though, it works only becuase of a continuou strategy space, I think.

FI' distinguishes between out-of-equilibrium moves and out-of-every-equilibrium moves, a distinction also made by the concept of "forward induction". Kohlberg (1989) says that forward induction requires players to make "deductions based on the opponents' rational behavior in the past" (p. 5), and that it is a special case of the principle that "a self-enforcing norm must be robust to the elimination of a strategy which is certain not to be employed where that norm is established" (p. 9). He discusses a variety of criteria that could go into the idea of forward induction, of which the most basic is:

THE FORWARD-INDUCTION REQUIREMENT (FI): *A self-enforcing outcome must remain self-enforcing when a strategy is deleted which is inferior (i.e., not a best reply) at every equilibrium with that outcome.*

FI and FI' both deal with what happens when an action is observed which is inferior at the current equilibrium. But making deductions and deleting inferior strategies are not necessarily the same idea. In some games FI and FI' reach the same conclusions, but in others they do not. In Joint Embezzlement they differ. FI does not refine perfect Bayesian equilibrium at all there, while FI' eliminates equilibria E_{2a} and E_{2b}, in which the outcome is $Fire$. E_2a and E_{2b} are the only equilibria with that outcome, and the boss's $Hire$ is not a best reply, but if $Hire$ is deleted from the game, it makes no difference to the equilibrium. FI therefore has no effect.

3.1. Three Players in the Game: Joint Embezzlement

A boss must decide whether to $Hire$ or $Fire$ two workers, Smith and Jones. Smith and Jones then simultaneously choose whether to $Work$ or $Steal$. Figure 1 shows the payoffs for the entire game in extensive form, and Table 1 shows the payoffs in the Smith-Jones subgame. If the boss chooses $Fire$, the payoffs are (0,0,0), and the strategies of Smith and Jones are irrelevant. If the boss chooses $Hire$, then whether Smith and Jones work or steal does matter. If both work, then output is high and the

workers receive wages, for payoffs of (3, 2, 2). If both steal, their theft is successful, for payoffs of $(-2, 4, 4)$. If Smith works and Jones steals, then output is moderate, Smith gets his wage plus a small bonus from turning in Jones, and Jones goes to jail, for payoffs of $(-1, 3, -6)$.

Figure 1: Joint Embezzlement.

		Jones	
		Work	Steal
Smith	Work	**2,2**	$3, -6$
	Steal	$-6, 3$	**4,4**

Payoffs to: (Smith, Jones).

Table 1: The Coordination Subgame from Joint Embezzlement.

An "equilibrium" is a strategy combination: one strategy for each player, chosen according to some rule favored by the modeller. The usual rule is that the strategy combination be a "perfect Bayesian equilibrium": the strategies are best responses to each other, the players follows Bayes's rule when possible, with beliefs specified by the modeller where Bayes' rule does not apply, and players' strategies must remain best responses regardless of the past history of the game. An "equilibrium outcome" is a path through the game tree generated by an equilibrium. There may be multiple equilibria, but only one can be played out in a given realization of the game; let us denote the equilibrium being played out as the "realized equilibrium".

The perfect Nash equilibria for Joint Embezzlement are:[9]

$E_1 : (Hire, Work, Work)$ with payoffs (3,2,2)

[9]There are also two non-perfect Nash equilibria:
$E_3 : (Fire, Work, Steal)$
$E_4 : (Fire, Steal, Work)$

E_{2a} : $(Fire, Steal, Steal)$ with payoffs (0,0,0)
E_{2b}:*(Fire, Work with probability 1/9, Work with probability 1/9)*

To reduce the number of equilibria further one needs to go beyond the generally accepted equilibrium concepts. The idea that will be modelled below is that E_1 should be the only equilibrium, because if the workers think that the realized equilibrium is E_{2a} or E_{2b} (which have the same outcome), but then observe the boss choosing $Hire$, each worker should become worried that the other worker is going to play according to what the boss seems to think is the realized equilibrium, E_1.[10]

Suppose it is common knowledge that E_2 is the realized equilibrium of Joint Embezzlement that is being played out. If the boss makes the off-equilibrium move of $Hire$, what is Smith to think? The boss choosing $Hire$ is a zero-probability event. One interpretation Smith might make is that the boss made a random mistake, in which case Smith's beliefs about the remainder of the game are unchanged. This is the response known as "passive conjectures" in signalling games. Alternatively, Smith might make a different interpretation: that he himself is confused and the equilibrium being played out by the other players is E_1, not E_2. If it is impossible for the boss to move accidentally, this is plausible, because the boss would be strictly worse off choosing $Hire$ if the realized equilibrium were E_2. The boss's move is a credible indicator of his confidence in E_1. If Smith believes this, he believes that Jones will pick $Work$, so $Work$ is Smith's own best response. Jones can either use the same reasoning or know that Smith is following it; either way, Jones will pick $Work$ too. If Smith and Jones behave according to this logic, the boss will certainly pick $Hire$ and E_1 is the only equilibrium that will ever be observed.

It seems here that out-of-equilibrium behavior changes a player's belief about which equilibrium is realized, which is possible only if he does

[10] This is a different problem from that of a boss with employees who collude to avoid a prisoner's dilemma subgame (e.g., Tirole, 1986). In Joint Embezzlement, the workers' problem is to coordinate on a Nash equilibrium, which is self-enforcing. Non-binding contracts between Smith and Jones might be important if there were no communication move, but if there is, then making the contract binding has no marginal effect. When Smith and Jones are in a tournament against each other, on the other hand, and they wish to collude, communication makes no difference. A non-binding contract would be useless, but making the contract binding would have a big marginal effect, since then both workers could trust each other to exert low effort.

not assign a prior of 100 percent to a single strategy combination being the realized equilibrium. A fundamental assumption behind Nash equilibrium is that the structure of the game and the identity of the realized equilibrium are common knowledge, an assumption which is violated if we specify that players put positive probability on more than one equilibrium being realized. Harsanyi (1967) showed that this assumption need not prevent the modeller from introducing asymmetric information about the structure of the game: the modeller simply adds a move at the start of the game in which Nature randomly chooses the game's structure, observed by some but not all players, using probabilities that are themselves common knowledge. Something similar can be done here to allow opinions to differ on which equilibrium is realized.

Let us follow Harsyani's approach of reformulating a conventionally unanalyzable game into one that can be analyzed using standard methods. First, let us construct an "extended game." The modeller begins by deciding which strategy combinations he considers to be candidates for equilibrium, using whatever criteria he finds plausible. Usually he will choose the set of perfect Bayesian equilibria, but other criteria might also be available; e.g., mixed strategies might be ruled out as implausible. Denote the outcomes of the strategies that survive this process as the set of *plausible equilibrium outcomes,* $E_i, i = 1, \ldots m$.

The *extended game* is the original game preceded by two moves by Nature. In the first move, Nature picks equilibrium outcome E_i with probability f_i, where $\sum_i f_i = 1$, unobserved by the players. In the second move, Nature sends a separate private message to each player j. With probability $(1 - \sum_k \epsilon_{ijk})$ the message is E_i, and with probability ϵ_{ijk} it is E_k, where $k \neq j$ and $\sum_k \epsilon_{ijk}$ is an arbitrarily small probability.[11]

ASSUMPTION 1: Player i believes that player j will follow the equilibrium revealed to j by Nature, X, until j discovers that some other player has deviated from X. This belief is common knowledge.

If any of the plausible equilibria fail to be subsets of perfect Bayesian equilibrium outcomes in the extended game, drop them and begin

[11]The natural extension to games with continua of equilibrium outcomes is to make f_i into a density $f(i)$ and ϵ_{ijk} into a function $\epsilon_j(ik)$ that integrates over i and k to an arbitrarily small probability.

again with the smaller revised set of plausible equilibria. Iterate this process until all m equilibrium outcomes are subsets of perfect Bayesian equilibrium outcomes of the extended game.

An FI' equilibrium is a strategy combination for the original game that generates an outcome which is a subset of a perfect Bayesian equilibrium outcome of the extended game once iteration has proceeded as far as possible.

The values of m, f_i and ϵ_{ijk} are chosen by the modeller. They represent the modeller's prior knowledge about the probability of different equilibrium outcomes. The particular values chosen affect the set of FI'-equilibria in some but not all games, as will be seen later. In the absence of special information, the natural values are the uniform ones: $f_i = 1/m$ and $\epsilon_{ijk} = \epsilon/(m-1)$ (so that $\sum_k \epsilon_{ijk} = \epsilon$). Unless otherwise specified, these uniform values will be used in the examples.

Since ϵ_{ijk} is arbitrarily small, it leaves behavior unaffected except for the potential to change players' beliefs about which equilibrium is realized. Its only effect is to allow Bayes' rule to interpret actions that would otherwise be out-of-equilibrium. Unlike the case where players makes mistakes because of "trembling hands," FI'-equilibrium does not allow Bayes' rule to interpret every possible action, only actions that are equilibrium actions for *some* plausible equilibrium.

Assumption 1 says that every player believes that the other players are deciding what is the realized equilibrium based on Nature's message. If a player receives message E_i, his belief is that E_i is going to be played by the other players unless they have received different messages. Since the probability they have received different messages is arbitrarily low, he will play X at least until he sees a deviation by another player. Thus, the behavior rules of the players under Assumption 1 are consistent with each other. Assumption 1 may seem novel, but it is in fact a weakened version of a standard but implicit assumption that Nash equilibrium requires for Nash behavior to be utility-maximizing:

ASSUMPTION 1′: Player i believes that player j will follow the equilibrium, X, revealed to j by Nature. This belief is common knowledge.

The initial moves by Nature together with Assumption 1 represent the background process by which the players arrive at common knowledge of the realized equilibrium. The process could be evolutive or eductive: it might be some kind of pre-game communication, or history, or psychological drives towards focal points, and it determines the probabilities f_i with which Nature chooses each equilibrium. Game theory has not yet determined this process, but Nash equilibrium implicitly assumes that it exists, and existence is all that we require. What is important is that somehow the process selects one strategy combination to be the realized equilibrium, and that the process contains a little noise. The ordinary concept of Nash equilibrium effectively relies on Assumption 1 or $1'$ and on Nature choosing $\epsilon = 0$, but this structure is concealed by using using a reduced form: the modeller just picks a strategy combination and asks whether any player would deviate unilaterally, without inquiring as to why that strategy combination was chosen. The approach I suggest brings this structure into the open, and what is new is the addition of a little noise.

This is not to be confused with the introduction of "cheap talk" into a game. The purpose of introducing Nature's moves is not to see what would happen if players could try to change the course of the game by costless communication. Nature's moves are not literal moves, but a heuristic to represent the background process by which the players choose which equilibrium they are going to play out. Allowing the players to ignore Nature's move, treating it as if it were "cheap talk" by a genuine player, would defeat the purpose of putting Nature's move into the game, since we would then require either (a) a new representation for the background process or (b) a return to the old assumption that the equilibrium being played out is common knowledge. It would be like saying that the players in a Bayesian game ought to be allowed to ignore the modeller when he tells them what priors they ought to hold. The expanded game is not so much adding assumptions to the game as replacing the assumption that the equilibrium being played out is common knowledge with the assumption that the background process represented by Nature's moves is common knowledge. What is truly new is not Assumption 1 and Nature's moves *per se*, but the possibility that Nature makes different announcements to different players.

FI'-equilibrium is thus based on the idea that if each player believes

that every other player follows the rule of playing out the current equilibrium ordained by the background process, he too should be willing to follow the rule. It rules out equilibria in which a player would profit by unilateral deviation. This would not reduce the number of equilibria in a game if the background process were perfectly coordinated, but if it is common knowledge that occasionally players come to different beliefs about which equilibrium is realized, then this has the potential to reduce the number of equilibria, as will be seen in the case of the game Joint Embezzlement.

To apply FI' to Joint Embezzlement, start with the two perfect equilibria E_1 and E_2.[12] Do they form a set of FI' equilibria? The first iteration of the extended game is:

(1) Nature chooses E_1 with probability 0.5 and E_2 with probability 0.5.

(2) Nature sends messages to the boss, Smith, and Jones. In each case the message is the equilibrium selected in move (1) with probability $(1 - \epsilon)$ and the unselected equilibrium with probability ϵ.

(3) The boss chooses $Hire$ or $Fire$.

(4) Smith and Jones simultaneously choose $Work$ or $Steal$.

E_1 is an FI' equilibrium because the boss has no incentive to deviate by choosing $Fire$ regardless of the effect of that deviation on beliefs. If he chooses $Fire$, that might induce the workers to choose (*Steal, Steal*) under the belief that the boss received the message "E_2" from Nature, but that deduction gives the boss no incentive to choose $Fire$.

E_2 is not an FI' equilibrium. Suppose that it were, that Nature had chosen E_2 to be the equilibrium, and that Nature sent the message E_2 to all three players without garbling. If the boss deviated by choosing $Hire$, how would Smith respond? Interpreting $Hire$ as equilibrium behavior, Smith would know that one of two things happened: Nature chose E_1 to be the equilibrium and sent the true message to the boss (and most likely to Jones) but sent Smith the garbled message E_2; or Nature chose E_2 to be the equilibrium but sent the boss the garbled message E_1. Either the boss

[12] Strictly speaking, E_2 is one equilibrium outcome, which results from the two equilibria E_{2a} and E_{2b}.

or Smith has been sent a garbled message, with equal probability. If the boss received a garbled message and E_2 is the equilibrium, then Smith should choose $Steal$, because it is almost certain that Jones will choose $Steal$, in which case $Steal$ yields a payoff of 4 and $Work$ yields only 3. If the boss received a correct message and E_1 is the equilibrium, then Smith should choose $Work$, because it is almost certain that Jones will choose $Work$, in which case $Work$ yields a payoff of 2 and $Steal$ yields -6. Given the asymmetric losses and the equal probabilities of mistakes by Smith and the boss, Smith should choose $Work$. The boss, foreseeing this, would choose $Hire$, so equilibrium E_2 is broken.

It was assumed above that the probabilities with which Nature selects each equilibrium are equal, and that the probabilities of garbled messages are the same for each player and for each message. If Nature selects E_1 with probability .1 and E_2 with probability .9, then E_2 *is* an equilibrium. In the extended game, if Smith observes $Hire$, he believes that there is a 90 percent probability that the Boss mistakenly received the message E_1, and only a 10 percent probability that he himself received a garbled message. In other games, Nature's probabilities do not matter.[13]

Whether this dependence on detail is an advantage or a disadvantage of FI' is open for debate. It explains why some examples are more compelling than others, an advantage, but it makes the idea context-dependent, so that it requires more thought to apply it. As an example in which FI' is particularly compelling and Nature's probabilities matter less, consider the following modification of Joint Embezzlement. Instead of one boss there are one hundred bosses, all partners in the business, and if even a single boss chooses $Fire$ instead of $Hire$, the payoffs are 0 for everyone. If the equilibrium is supposed to be E_2, but one hundred bosses in a row do the unthinkable and choose $Hire$, what is Smith to believe? An equilibrium like E_2, in which Smith responds by choosing $Steal$, seems unreasonable. Under FI', Smith would believe that it is he who received the garbled message, rather than every one of the

[13] An example is the extended version of the PhD Game in Section 4: the professor would accept the applying student regardless of the probabilities of E_1 and E_2 and of garbling. Those probabilities are irrelevant because it is not Nature's choice of equilibrium that matters to the professor, but Nature's message to the student. Even if the professor knows that it is the student, and not himself, who received the garbled message, the professor will go along with the equilibrium suggested by the student's behavior.

hundred bosses, and this would be true even if E_1 has a very low probability of being selected by Nature. Thus, FI' can differentiate between the game with one boss and the game with one hundred.[14]

[14]FI, discussed below, cannot make this distinction.

4. Incomplete Information and Beliefs about Types

Until recently, equilibrium refinements were discussed almost solely in the context of games of incomplete information, in which Nature makes an initial move and chooses the "types" of the players. Each player has a prior belief on the types of the other players that he updates into a posterior belief as events occur, using Bayes' rule and the out-of-beliefs specified by the particular perfect Bayesian equilibrium. Many refinements place restrictions on the out-of-equilibrium beliefs. This makes it seem as if the refinement is essentially about updating beliefs on types, but it is not possible to restrict beliefs on types without restricting beliefs on which equilibria are being played out. The examples earlier in this paper show that the issue of beliefs about which equilibrium is being played out arise whether information is asymmetric or not, and suggest that the emphasis on updating beliefs about types clouds the basic issue.

But it is desirable to have an equilibrium refinement that can help determine the equilibrium in games of both complete and incomplete information. This section turns to three games of incomplete information to show the effect of FI'. The first game, "The PhD Game," has both separating and pooling perfect bayesian equilibria. FI' eliminates the pooling equilibrium in much the same way as the intuitive criterion does. The second game, "The Beer-Quiche Game," has two pooling equilibria, one preferred by each type of player. In this game, the intuitive criterion eliminates one equilibrium, but FI' does not. The third game, a signalling game with a continuum of actions, has a continuum of separating and a continuum of pooling equilibria. FI' can reduce these to a single equilibrium.

Consider first "the PhD Game" from Rasmusen (1989), in which a student who is either smart or stupid must decide whether to apply to graduate school and a professor must decide whether to admit students who apply.

(0) Nature chooses the type of the student to be either Smart (probability 0.1) or Stupid (probability 0.9), unobserved by the professor.

(1) The student decides to *Apply* to graduate school, at some cost, or *Not Apply.*

(2) If the student chose *Apply*, the professor decides whether to *Accept* or *Reject* him.

		Professor	
		Accept	Reject
Applying Student	Smart	10, 10	$-1, 0$
	Stupid	$-10, -10$	-1,0

Payoffs to: (Student, Professor).
If the Student chooses Not Apply, payoffs are (0,0).

Table 4: The PhD Game

The smart student would like to be accepted, and the professor would like to accept him. For the stupid student it is a strictly dominant strategy not to apply; he actually lowers his payoff by successfully pretending to be smart and being admitted.

The game has two perfect Bayesian equilibrium outcomes. In the separating equilibrium, E_1, the smart student chooses *Apply*, the stupid student chooses *Not Apply*, and the professor accepts anyone who applies. In the pooling equilibrium, E_2, neither type of student applies, and the professor would reject anyone who did choose *Apply*, under any out-of-equilibrium belief that specifies that the probability that a mistaken application is by a stupid student is over 0.5.

E_2 is perverse, but its status as an equilibrium under Nash logic is impeccable: no player has any incentive to deviate, and the result is not due to a very contrived choice of out-of-equilibrium beliefs. Passive conjectures, for example, would lead the professor to put a probability of 0.9 that a student who deviates and applies is stupid.

E_2 is not, however, an FI'-equilibrium, because in the extended game a student who chooses *Apply* might be doing so under the belief that the realized equilibrium is E_1. The first iteration of the extended game, which treats both E_1 and E_2 as plausible, is:

(A1) Nature chooses E_1 with probability 0.5 and E_2 with probability 0.5.

(A2) Nature sends messages to the student and the professor. In each case the message is the equilibrium selected in move (A1) with probability $(1 - \epsilon)$ and the unselected equilibrium with probability ϵ.

(0) Nature chooses the type of the student to be either Smart (probability 0.1) or Stupid (probability 0.9), unobserved by the professor.

(1) The student decides to *Apply* to graduate school, at some cost, or *Not Apply*.

(2) If the student chose *Apply*, the professor decides whether to *Accept* or *Reject* him.

If the student chooses *Apply*, the professor deduces that the student heard E_1 from Nature and is acting accordingly—in which case the move of *Apply* is a sure sign that the student is smart. Note too that that this result is independent of the probabilities used in moves (A1) and (A2). E_2 is not an FI'-equilibrium, because in any specification of the extended game that includes both E_1 and E_2, a student who chooses *Apply* might be doing so under the belief that the realized equilibrium is E_1.

This seems very different from the usual approach to refining equilibrium, which is based on restricting the out-of-equilibrium beliefs that might be held. Possibly the best-known refinement is the "intuitive criterion" of Cho & Kreps (1987), which is defined for games in which the first mover is trying to communicate private information about his type, t'. The intuitive criterion says that if he takes an out-of-equilibrium action m' which could not possibly be profitable were he of type t, the other players should respond by assigning zero probability to type t. In the words of Cho and Kreps (1987, p. 181), if the first player sends the following message he

should be believed:

> "I am sending the message m', which ought to convince you that I know t'. For I would never wish to send m' if I know t, while if I know t', and if sending this message so convinces you, then, as you can see, it is in my interest to send it."

It might seem that this is quite different from FI'—that the Cho-Kreps player is trying to communicate private information, whereas under FI' the first player is trying to change later players' beliefs as to which equilibrium is realized. But the difference is superficial. The Cho-Kreps player is not just trying to communicate his private information: when he sends a message impossible in a particular equilibrium, he is trying to convince the other players to change their beliefs *about which equilibrium is realized* in order to use their new belief about the equilibrium to convey a new belief about his type. If he fails to convince other players about the equilibrium, he will not change their beliefs about his type.

In the PhD Game, the intuitive criterion rules out E_2 by drastically restricting beliefs. It notes that the stupid student would have no incentive to *Apply* even if this would change the professor's out-of-equilibrium belief to one that would induce him to *Accept*, and concludes that the professor's out-of-equilibrium belief should be that an applying student is smart, which eliminates E_2 as an equilibrium.[15]

The intuitive criterion and FI' reach the same result in this example, but to accept the intuitive criterion one must believe that some beliefs are self-evidently reasonable or that equilibrium-dominated strategies are axiomatically disqualified. Under FI', on the other hand, the out-of-equilibrium belief is the inevitable consequence of a less-than-perfect process for deciding which equilibrium is being played out.

[15]FI also rules out E_2, on the more formal ground that if the equilibrium-dominated strategy of (*Apply* if stupid) is dropped, then a deviation to (*Apply* if smart) becomes profitable.

A second example, the Beer-Quiche Game of Cho & Kreps (1986), will show that in some games of incomplete information, FI' cannot reduce the number of equilibria at all, even if one of the equilibria is based on out-of-equilibrium beliefs that seem very unreasonable.

In this game, Player I might be either weak or strong. Player II wishes to fight a duel only if Player I is weak, which has a probability of .1. Player II also observes what Player I has for breakfast, and he knows that weak players prefer quiche for breakast, while strong players prefer beer. Player I wishes above all to avoid a duel, regardless of his type. The payoffs are as shown in Figure 4.

Figure 2: The Beer-Quiche Game. . (see Kreps article fgiure 2.13).

This game has two perfect Bayesian equilibrium outcomes, both of which are pooling. In E_1, Player I has beer for breakfast regardless of type, and Player II chooses not to duel. This is supported by the out-of-equilibrium beliefs that a quiche-eating Player I is weak with probability over 0.5, in which case Player II would choose to duel on observing quiche. In E_2, Player I has quiche for breakfast regardless of type, and Player II chooses not to duel. This is supported by the out-of-equilibrium beliefs that a quiche-eating Player I is weak with probability over 0.5, in which case Player II would choose to duel on observing quiche.

The intuitive criterion rules out E_2, on the grounds that if Player I deviates and says "I am having beer for breakfast, which ought to convince you that I am strong. For I would never wish to have beer for breakfast if I were weak, while if I am strong, and if sending this message so convinces you, then, as you can see, is in my interest to have beer for breakfast."

FI' cannot rule out E_2. If the plausible equilibrium outcomes are E_1 and E_2, then in the extended game Nature chooses both of these with positive probabilities whose exact values do not matter. Suppose that Players I and II have both received message E_1 and Player I is weak, so that E_1 tells him to choose beer. If Player I deviates and chooses quiche, then Player II's conclusion will be that Player I thinks the equilibrium is E_2, and that Player I is strong with probability 0.9. Player II therefore will not duel. This would make deviation profitable and rule out E_1, leaving E_2 as the equilibrium. Suppose, on the other hand, Players I and II have both received message E_2 and Player I is strong, so that E_2 tells him to choose beer. If Player I deviates and chooses beer, then Player II's conclusion will be that Player I thinks the equilibrium is E_1, and that Player I is strong with probability 0.9. Player II therefore will not duel. This would make deviation profitable and rule out E_2, leaving E_1 as the equilibrium. Thus, by choosing the order of iteration, either E_1 or E_2 can be made to survive as FI' equilibria.

There is an approach similar in spirit to FI' which can rule out E_2. Suppose that we construct an extended game in which Nature chooses not between two realized equilibria, but between two games. In Game X, Player II observes what Player I has for breakfast; in Game Y, Player II does not. Nature tells each player which game is being played, but with some small probability the game is X and Nature tells Player I that the game is Y. In this case, Player I will have beer or quiche for breakfast, depending on which breakfast he prefers. Knowing that this is a possibility, Player II would interpret out-of-equilibrium quiche as indicating a weak Player I and out-of-equilibrium beer as indicating a strong Player I. This in turn both supports the out-of-equilibrium beliefs necessary for E_1 and rules out the beliefs necessary for E_2.

A third example will be presented to illustrate the application of FI' to a game of incomplete information with a continuum of perfect Bayesian equilibria, to show how these can be reduced to a single FI' equilibrium. Let there be two types of workers, of ability $a = 1$ and $a = 2$, where 0.8 are of type 1 and 0.1 are of type 2. Workers of type i acquire s_i years of education at cost s_i/a, where s_i is a continuous non-negative variable. Employers will pay a worker according to their estimate of his ability, but they observe only education.

This game has a continuum of separating perfect Bayesian equilibrium outcomes, in which $s_1^* = 0$ and $s_2^* \in [1, 2]$, and $w(s = 0) = 1$ and $w(s = s_2^*) = 2$. An out-of-equilibrium belief that supports such an outcome is that any worker who acquires an education level not specified by the particular equilibrium is of type 1.[16] The type-1 workers will wish not to acquire education because the benefit is a salary increase of 2-1, but the cost is at least 1/1. The type-2 workers will wish to acquire education because the benefit is 2-1 and the cost is no more than 2/2.

This game also has a continuum of pooling perfect Bayesian equilibrium outcomes, in which $s_1 = s_2 \in [0, 0.2]$ and $w(s = s^*) = 1.2$. An out-of-equilibrium belief that supports such an outcome is that any worker who acquires an education level not specified by the particular equilibrium is of type 1 (again, there is a continuum of equilibria supporting each equilibrium outcome) Any deviating worker would deviate to $s = 0$. The type-1 workers will acquire education because the benefit is a salary increase of 1.2-1, and the cost is no more than .2/1. The benefit is the same for type-2 workers, and the cost is even less.[17]

The initial extended game is:

(1) Nature picks an equilibrium outcome. With probability 1/6 it is a pooling equilibrium and s_2^* takes values in $[0, 0.2]$ with uniform density. With probability 5/6 it is a separating equilibrium and s_2^* takes values in $[1, 2]$ with uniform density.

(2) Nature announces an equilibrium outcome to each player. With probability $(1 - \epsilon)$ it is the outcome chosen in (1). With probability ϵ it is not, and the announcement is drawn randomly using the same distributions as in (1).[18]

(3) Nature chooses the worker to have ability $a = 1$ with probability 0.8 and $a = 2$ with probability 0.2.

(3) The worker chooses education s.

[16] There is a continuum of equilibria that support each outcome, differing in the out-of-equilibrium beliefs; e.g., if $s_2^* = 1.5$ replace Prob(type $1|s_2 = 0.3) = 1$ with Prob(type $1|s_2 = 0.3) = .99$.

[17] xxx There are probbaly semi-pooling euqilibria too; I haven't thoght about it.

[18] Note that the probability of drawing the same outcome twice from a continuous distribution is zero.

(4) Two employers compete for the services of the worker by simultaneously offering him wages $w(e_1)$ and $w(e_2)$.

Suppose first that the message from Nature is the most attractive pooling equilibrium, with $s^* = 0$, but the worker picks $s = 1$. The employer will conclude that the worker is type 2 and thinks the realized equilibrium is separating. The worker will be paid 2 instead of 1.2, and he will incur a cost of 1/2, so if he is indeed type-2 he will benefit from this. Type-2 workers will deviate from even the most attractive pooling equilibrium, ruling out that class of equilibria. But we can also rule out every separating equilibrium but $s_2^* = 1$, because the employer will interpret a deviation to $s_2 = 1$ as the playing out of an $s_2^* = 1$ equilibrium, which leaves the type-2 worker with the same wage but a lower signalling cost.

5. Concluding Comments

My intent in this article has not been to persuade the reader to adopt a new axiom that he must apply to the equilibrium of every game. Rather, I hope to help him recognize the implications of something he very likely accepts already: Nash equilibrium. If one accepts Nash equilibrium when there are multiple equilibria, then one accepts the existence of a process by which the players are apprised of the equilibrium to be played out. If one furthermore accepts that the process is imperfect, then one has accepted FI' or something very like it.

The simplest way to describe FI' is as a requirement that players in a game interpret actions as equilibrium actions whenever possible, even if they must change their beliefs about what equilibrium is being played out. Most refinements of equilibrium suggested in recent years are based on what out-of-equilibrium beliefs seem intuitively reasonable. Often the idea of weak dominance is invoked— that a strategy should work well even if the world is not exactly as the player thinks it is. FI' tries to provide a more fundamental grounding for our intuition. It traces what seems intuitively reasonable to the idea that when there are multiple equilibria the identity of the equilibrium being played is only almost common knowledge and the players are aware that their beliefs might be inconsistent with those of other players. FI' reduces the number of equilibria in a way very similar to forward induction. This suggests that the idea behind FI' is also a justification for forward induction, and, since the two concepts do not always lead to exactly the same predicted equilibria, it suggests an improvement upon forward induction as well.

The reader of a theory article sometimes turns the last page with the feeling that the conclusions are all very well, but no application for them ever existed or ever will exist. I have shown that FI' and FI differ in a certain kind of three-player dynamic game. Are there any useful games of this kind? The situation does not seem unusual: a dominant player sets the agenda and a number of other players wish to pick the same response from a number of possibilities. The specific example which inspired this article is the model of exclusionary practices in Rasmusen, Ramseyer, and Wiley (1990). In that model, a firm may require its customers to sign exclusive-dealing contracts. If most customers sign, other firms drop out of the market, so the individual is no worse off than if he had not signed. Thus, there is a coordination game among customers, who would jointly prefer not to sign the contract, but who will sign in exchange for a small payment from the firm if each believes that the others will sign. The game has two perfect equilibria, one in which the contract is offered and signed, and one in which it is not offered. The reasoning of FI' says that if the firm, at some cost, offers the contracts, each customer must consider whether the firm has better knowledge of the equilibrium than he does. In this example, the customers, with less at stake, would reasonably have a greater chance than the firm of being confused over the equilibrium being played out. In that case, the no-exclusion equilibrium disappears, which suggests a reason for firms to use and anti-trust authorities to worry about exclusion contracts. Thus, the choice of equilibrium may matter to profitability and policy.

References.

Ben-Porath, Elchanan & Eddie Dekel (unpublished) "Coordination and the Potential for Self Sacrifice" working paper, Dept of Economics, U. of Cal, Berkeley, May 1989.

Binmore, Ken (1987) "Modelling Rational Players, Part I" *Economics and Philosophy.* October 1987. 3, 2: 179-214.

Binmore, Ken (1990) *Essays on the Foundations of Game Theory.* Oxford: Basil Blackwell Ltd., 1990.

Carlsson, Hans & Eric van Damme (1994) "Global Games and Equilibrium Selection Hans Carlsson," *Econometrica*, 61(5): 989-1018 (September 1993).

Cho, In-Koo & David Kreps (1987) "Signalling Games and Stable Equilibria" *Quarterly Journal of Economics.* May 1987. 102, 2: 179-221.

Farrell, Joseph (1987) "Cheap Talk, Coordination, and Entry" *Rand Journal of Economics.* Spring 1987. 18, 1: 34-9.

Harsanyi, John (1967) "Games with Incomplete Information Played by 'Bayesian' Players, I: The Basic Model" *Management Science.* November 1967. 14, 3: 159-82.

Kohlberg, Elon (unpublished) "Refinement of Nash Equilibrium: The Main Ideas," Harvard Business School working paper #89-073, June 1989.

Kohlberg, Elon & Jean-Francois Mertens (1986) "On the Strategic Stability of Equilibria" *Econometrica.* September 1986. 54, 5: 1003-7.

Morris, Stephen & Hyun Song Shin "Global Games: Theory and Applications,"

Myerson, Roger (1978) "Refinements of the Nash Equilibrium Concept" *International Journal of Game Theory.* 1978. 7, 2: 73-80.

Okuno-Fujiwara, Masahiro & Andrew Postlewaite (unpublished), "Forward Induction and Equilibrium Refinement," CARESS working paper #87-01, February 1987.

Osborne, Martin (unpublished), "Signaling, Forward Induction, and Stability in Finitely Repeated Games," working paper, Dept of Economics, McMaster University, November 1987.

Rasmusen, Eric (1989) *Games and Information*. Oxford: Basil Blackwell, 1989.

Rasmusen, Eric, Mark Ramseyer & John Wiley (unpublished) "Naked Exclusion" UCLA AGSM Business Economics Working Paper #89-17.

Tirole, Jean (1986) "Hierarchies and Bureaucracies. On the Role of Collusion in Organizations." *Journal of Law, Economics, and Organization"* Vol 2, no. 2, Fall 1986, pp. 181-214.

Van Damme, Eric (1987) *Stability and Perfection of Nash Equilibrium.* Berlin: Springer-Verlag, 1987.

Van Damme, Eric (1989) "Stable Equilibria and Forward Induction," *Journal of Economic Theory*, 48: 476-496.