Aspects of Rationalizable Behaviour

Peter J. Hammond

Department of Economics, Stanford University, CA 94305-6072, U.S.A.

December 1991; revised April, July, and September 1992. Chapter for *Frontiers of Game Theory* edited by Ken Binmore, Alan Kirman, and Piero Tani (M.I.T. Press).

ABSTRACT

Equilibria in games involve common "rational" expectations, which are supposed to be endogenous. Apart from being more plausible, and requiring less common knowledge, rationalizable strategies may be better able than equilibria to capture the essential intuition behind both correlated strategies and forward induction. A version of Pearce's "cautious" rationalizability allowing correlation between other players' strategies is, moreover, equivalent to an iterated procedure for removing all strictly, and some weakly dominated strategies. Finally, as the effect of forward induction in subgames helps to show, the usual description of a normal form game may be seriously inadequate, since other considerations may render implausible some otherwise rationalizable strategies.

ACKNOWLEDGEMENTS

I am grateful to Luigi Campiglio of the Università del Sacro Cuore in Milan and to Piero Tani of Florence for making possible the lectures at which some of these ideas were originally presented. In addition, Piero Tani coaxed me into giving a talk at the conference, provided me time to write up the ideas presented, and has been remarkably good-humoured throughout the whole process. A useful interchange with Giacomo Costa of the Università di Pisa did much to convince me of the interest of forward induction. Pierpaolo Battigalli kindly pointed out a serious error in earlier versions, and passed on more good suggestions than I have so far been able to explore.

1. Introduction

1.1. Common Expectations Equilibrium

For much of its history, game theory has been pursuing an enormously ambitious research program. Its aim has been nothing less than to describe equilibrium outcomes for every game. This involves not only a specification of each player's behaviour or choice of strategy, but also of what beliefs or expectations justify that behaviour. The crucial feature of equilibrium expectations is that they correspond to a single *common* joint probability distribution over all players' strategy choices. This makes it natural to speak of *common expectations equilibrium*. Moreover, these expectations must be *rational* in the sense that each player is believed to play, with probability one, some optimal strategy given the conditional joint probability of all the other players' strategies. Indeed, this is the essential idea behind Aumann's (1987) notion of correlated equilibrium. If one also imposes stochastic independence between the probability beliefs concerning strategy choices by different players, then each common expectations or correlated equilibrium with this property is a Nash equilibrium, and conversely. Thus, the crucial feature of both Nash and correlated equilibrium is the existence of common "rational" expectations. In the case of games of imperfect information, the same is true of a Bayesian-Nash equilibrium with a common prior, as considered by Harsanyi (1967–8). Refinements of Nash equilibrium, including subgame perfect, trembling hand perfect, proper, sequential, and stable equilibria, all incorporate the same common expectations hypothesis *a fortiori*.

1.2. Battle of the Sexes, and Holmes v. Moriarty

		C	
		a	*b*
R	*A*	2 1	0 0
	B	0 0	1 2

Battle of the Sexes

The restrictiveness of such common expectations equilibrium can be illustrated by two classic games. The first is "Battle of the Sexes", as presented by Luce and Raiffa (1957) and often considered since. As is well known, this game between the two players Row (R) and Column (C) has three Nash equilibria in which:

(1) Row and Column play (A, a) and get $(2, 1)$;

(2) Row and Column play (B, b) and get $(1, 2)$;

(3) Row and Column play the mixed strategies $(\frac{2}{3}\,A + \frac{1}{3}\,B, \frac{1}{3}\,a + \frac{2}{3}\,b)$, and get expected payoffs $(\frac{2}{3}, \frac{2}{3})$.

Knowing that these are the three Nash equilibria does little to help select among them. Moreover, while observing the strategy choices (A, a) or (B, b) might help to reinforce our belief in equilibrium theory, how are we to interpret an observation of (B, a) or of (A, b)? Did these anomalous observations arise because the players followed the mixed strategies of the third Nash equilibrium and experience misfortune on this occasion? Or if (A, b) is observed, is that because R hoped that the equilibrium would be (A, a) whereas C hoped that it would be (B, b)? And if (B, a) is observed, is that because R feared that the equilibrium would be (B, b) while C feared it would be (A, a)?

A second classic game is von Neumann and Morgenstern's (1953, pp. 176–8) simple model of one small part of the story concerning Sherlock Holmes' attempts to escape being murdered by Professor Moriarty before he could get the professor and his criminal associates arrested.[1] According to this simple model Moriarty wants to catch Holmes, while Holmes wants to escape to Dover and then to the Continent. As Holmes boards a train heading from London to Dover, he sees Moriarty on the platform and sees that Moriarty sees him, and sees him seeing Moriarty, etc.[2] Holmes knows that Moriarty will hire a special train to overtake him. The regular train has just one stop before Dover, in Canterbury. Given the payoffs which von Neumann and Morgenstern chose in order to represent Holmes' and Moriarty's preferences, this involves the two person zero-sum bimatrix game shown. Here

[1] I am indebted to Kenneth Arrow for reminding me of this example which I had forgotten long ago. For the full story, including Holmes' finding himself in Florence sometime in May 1891, see "The Final Problem" in Sir Arthur Conan Doyle's *Memoirs of Sherlock Holmes*, followed by "The Adventure of the Empty House" in the same author's *The Return of Sherlock Holmes*. Morgenstern seems to have been fond of this example, since he had already used it twice before — see Morgenstern (1928, p. 98, and 1935, p. 343) [and note that the reference to the latter in von Neumann and Morgenstern (1953, p. 176) is slightly inaccurate].

[2] Actually, in "The Final Problem" it is not at all clear that Moriarty does see Holmes, who had after all taken the precaution of disguising himself as "a venerable Italian priest" so effectively as to fool his constant companion Dr. Watson. Nevertheless, Moriarty acted as though he knew Holmes really was on the train, and he would surely expect to be seen by Holmes.

		M	
		c	*d*
H	*C*	-100 100	0 0
	D	50 -50	-100 100

Holmes v. Moriarty

H denotes Holmes, *M* denotes Moriarty, and each has the strategy *D* or *d* of going on to Dover, as well as the strategy *C* or *c* of stopping in Canterbury.

The unique equilibrium of this simplified game has Moriarty stop his special train in Canterbury with probability 0.4, but go on to Dover with probability 0.6, while Holmes should get off the ordinary train in Canterbury with probability 0.6 and go directly to Dover with probability 0.4. In fact, Conan Doyle had Holmes (and Watson) get out at Canterbury and take hiding as Moriarty's special train rushed through the station on its way to Dover. Von Neumann and Morgenstern observe that this is the closest pure strategy profile to their mixed strategy Nash equilibrium, but that the latter gives Holmes only a 0.52 chance of escaping. In fact, who is to say that Holmes did not do better? He predicted that Moriarty would head straight to Dover, thinking perhaps that he had to reach it at all costs in time to intercept Holmes before he could escape quickly, while believing (correctly) that he could still catch Holmes later even if Holmes did get off at Canterbury and proceed more slowly to the Continent. What this shows, I suppose, is partly that von Neumann and Morgenstern's model is excessively simplified. But it also calls into question their advocacy of mixed strategies in such complex situations. Though pursuit games can involve bluff, just as poker does, there is much else to consider as well.

In the first of these two games there is as yet no good story of how one of the three Nash equilibria is to be reached. In the second, even though there is a single Nash equilibrium in mixed strategies, it is by no means obvious that this describes the only possible set of rational beliefs which the players might hold about each other.

1.3. Rationalizable Strategic Behaviour

Fortunately, work on "rationalizable strategic behaviour" initiated during the early 1980's by Bernheim (1984, 1986) and Pearce (1984) (and in their earlier respective Ph. D. theses) offers us a way out of the predicament. Actually, their approach is quite closely related to Farquharson's (1969, ch. 8 and Appendix II) earlier idea (in his 1958 D. Phil. thesis) that one should eliminate all (weakly) dominated strategies repeatedly, and consider anything left over as a possible strategy choice. A partial link between these ideas will be discussed in Section 3 below. In both Battle of the Sexes and Morgenstern's version of Holmes v. Moriarty, *any* pair of strategy choices by both players is rationalizable. However, before we can say what is really rational, we have to say much more about players' beliefs, and about whether they are sensible. This seems entirely right. A Nash equilibrium is indeed plausible when there is good reason for the players to have the particular common expectations underlying that equilibrium. But if there is no particularly good reason for them to have such common expectations, a Nash equilibrium has no more obvious claim to our attention than does any other profile of rationalizable strategies with associated divergent beliefs for the different players.

1.4. Outline of Paper

The rest of this paper will study some implications of relaxing the hypothesis that expectations are held in common by all players, and so of allowing that any profile of rationalizable strategies could occur. Section 2 begins by considering iteratively undominated strategies. Thereafter, it provides a concise definition of rationalizable strategies, and explains how they form a subset of those strategies that survive iterative deletion of strictly dominated strategies. An example shows that this rationalizability is too weak a criterion, however, because it allows players to use strategies which are strictly dominated in subgames, but not in the whole game. Accordingly, Pearce's more refined concept of "cautious rationalizability" is recapitulated for later use. It is also shown how strategies are cautiously rationalizable only if they survive a "cautious" version of iterative deletion of weakly dominated strategies.

Thereafter, Section 3 will consider correlated strategies, and argue that there is more scope for these in connection with divergent expectations than with the common expectations that underlie a correlated equilibrium. It is somewhat implausible for players to believe, even in the absence of a correlation device, that their own strategies are correlated with those of other players. But it is entirely reasonable for one player to believe that strategies of other players are correlated with each other. The common expectations hypothesis forces these two kinds of correlation to be synonymous, but in games with three or more players the ideas are quite different when players' expectations are allowed to diverge. Moroever, the set of all correlated rationalizable strategies can be found by removing all strictly dominated strategies for each player iteratively. And the set of all correlated cautiously rationalizable strategies can be found by "cautiously" iterating the rule of removing weakly dominated strategies.

Though rationalizability appears to be a coarsening instead of a refinement of Nash equilibria, there are cases when it helps to establish a refined equilibrium, as both Bernheim (1984, p. 1023) and Pearce (1984, p. 1044) pointed out. Indeed, as Section 4 argues, the arguments behind forward induction make much more sense when players are allowed to be uncertain about what happens in a subgame, instead of having in mind some definite Nash or sequential equilibrium. The arguments that help sustain forward induction, however, also point to a serious weakness of orthodox game theory. For it turns out that players' expectations in a subgame may well be influenced by the opportunities which they know other players have foregone in order to reach the subgame. Indeed, such is the essence of forward induction. But then it follows that the subgame is not adequately described by the strategy sets and payoff functions, or even by its extensive form tree structure. Important information about what happened before the subgame started is missing from the usual description of a game, and such information could also be relevant to the whole game, since that is presumably a subgame of some larger game which started earlier.

Iterative removal of all weakly dominated strategies has often been criticized as leading to implausible outcomes. Some of these objections are based on examples due to van Damme (1989) and also to Ben-Porath and Dekel (1992) of how the opportunity to "burn money" at the start of a game can significantly influence its outcome, even though that opportunity is never actually used. Dekel and Fudenberg (1990) discuss another similar example. Though cautious rationalizability is generally different from iterative removal of all weakly dominated strategies, in these particular examples it actually leads to identical outcomes. Accordingly, Section 4 also provides a brief discussion of these examples. It is claimed that appropriate forward induction arguments make the disputed outcomes less implausible than has previously been suggested.

Section 5 offers some brief concluding remarks.

2. Iteratively Undominated and Rationalizable Strategies

2.1. Iteratively Undominated Strategies

Consider a normal form game (N, A^N, v^N), where N is a finite set of players who each have specified finite (action) strategy spaces A_i, and A^N denotes the Cartesian product space $\prod_{i \in N} A_i$ of strategy profiles, while $v^N = (v_i)_{i \in N}$ is a list of all individuals' payoff functions $v_i : A^N \to \Re$ which depend on the strategy profile $a^N \in A^N$.

As general notation, given any measurable set T, let $\Delta(T)$ denote the set of all possible probability distributions over T.

Given any product set $K^N = \prod_{i \in N} K_i \subset A^N$ of strategy profiles, any player $i \in N$, and any pure strategy $a_i' \in K_i$, say that a_i' is *strictly dominated relative to* K^N if there exists $\mu_i \in \Delta(K_i)$ such that $\sum_{a_i \in K_i} \mu_i(a_i)\, v_i(a_i, a_{-i}) > v_i(a_i', a_{-i})$ for all combinations $a_{-i} = (a_j)_{j \in N \setminus \{i\}} \in K_{-i} := \prod_{j \in N \setminus \{i\}} K_j$ of all the other players' pure strategies. And say that a_i' is *weakly dominated relative to* K^N if there exists $\mu_i \in \Delta(K_i)$ such that $\sum_{a_i \in K_i} \mu_i(a_i)\, v_i(a_i, a_{-i}) \geq v_i(a_i', a_{-i})$ for all $a_{-i} \in K_{-i}$, with strict inequality for at least one such a_{-i}. Then let $S_i(K^N)$ (resp. $W_i(K^N)$) denote the members of K_i that are not strictly (resp. weakly) dominated relative to K^N. Also, let $S(K^N)$ and $W(K^N)$ denote the respective product sets $\prod_{i \in N} S_i(K^N)$ and $\prod_{i \in N} W_i(K^N)$.

Next, for each positive integer $k = 1, 2, 3, \ldots$, let $S^k(K^N) := S(S^{k-1}(K^N))$ be defined recursively, starting from $S^0(K^N) := K^N$. In the case when $K^N = A^N$, write S^k for $S^k(A^N)$, and let S_i^k denote player i's component of the product space $S^k = \prod_{i \in N} S_i^k$; it is the set of all i's strategies that remain after k rounds of removing all the strictly dominated strategies of every player. Evidently

$$\emptyset \neq S_i^k \subset S_i^{k-1} \subset \ldots \subset S_i^1 \subset S_i^0 = A_i \quad (k = 3, 4, \ldots).$$

Therefore, because each player's A_i is a finite set, the limit set

$$S_i^\infty := \lim_{k\to\infty} S_i^k = \bigcap_{k=0}^{\infty} S_i^k$$

must be well defined and non-empty. Actually, there must exist some finite n for which $S_i^k = S_i^n$ for all $i \in N$ and all integers $k \geq n$. Moreover, the same arguments apply to the sets W_i^k of strategies that survive k rounds of removing all the weakly dominated strategies of every player, and for the limit sets W_i^∞.

Finally, to prepare the ground for the later discussion of cautious rationalizability, another iterative rule for removing dominated strategies needs to be considered. This involves the recursive definition that starts from $(W\,S^\infty)^0(A^N) := A^N$ and continues with

$$(W\,S^\infty)^k(A^N) := W(S^\infty((W\,S^\infty)^{k-1}(A^N))) \quad (k = 1, 2, 3, \ldots),$$

where $S^\infty(K^N) := \bigcap_{k=0}^{\infty} S^k(K^N)$, of course. Also, $(W\,S^\infty)^k(A^N)$ can be written as the Cartesian product $\prod_{i\in N} (W\,S^\infty)_i^k$. Arguing as before, this recursion converges in a finite number of steps, so that each player $i \in N$ has a well defined non-empty limit set

$$(W\,S^\infty)_i^\infty := \bigcap_{k=0}^{\infty} (W\,S^\infty)_i^k.$$

In what follows, $(W\,S^\infty)_i^\infty$ will be described as the set of i's strategies that survive *cautious iterative deletion of dominated strategies.* Evidently, the construction of this limit set begins with removing strictly dominated strategies iteratively, in order to arrive at S_i^∞ for each player. But whenever the process of removing strictly dominated strategies gets stuck because there are no more to remove, the procedure passes on to the next stage which consists of just *one* round of removing *all* the weakly dominated strategies of *every* player. After this single round, it reverts immediately to removing strictly dominated strategies iteratively. Since the process can get stuck many times, in fact strategies that are only weakly dominated may have to be removed repeatedly. But such weakly dominated strategies are removed only "cautiously," in the sense that they must remain inferior even after strictly dominated strategies have been removed iteratively as far as possible.

2.2. Definition and Basic Properties of Rationalizable Strategies

For any probability distribution $\pi_i \in \Delta(A_{-i})$ which represents player i's beliefs about other players' strategies $a_{-i} \in A_{-i}$, let

$$U_i(a_i, \pi_i) := \mathbb{E}_{\pi_i} v_i(a_i, a_{-i}) := \sum\nolimits_{a_{-i} \in A_{-i}} \pi_i(a_{-i})\, v_i(a_i, a_{-i})$$

denote, for each $a_i \in A_i$, the expected value of v_i with respect to π_i. Then i's *best response correspondence* is defined by

$$B_i(\pi_i) := \arg\max_{a_i \in A_i} \{\, U_i(a_i, \pi_i) \,\}.$$

The sets R_i $(i \in N)$ of *rationalizable strategies* are now constructed recursively as follows (see Pearce, 1984, p. 1032, Definition 1, and compare Bernheim, 1984, p. 1015). First, let $R_i^0 := A_i$ for each player i. Then let

$$R_i^k := B_i \left(\prod\nolimits_{j \in N \setminus \{i\}} \Delta(R_j^{k-1}) \right) \quad (i \in N;\ k = 1, 2, \ldots).$$

Thus R_i^1 consists of all player i's possible best responses, given the various beliefs that i might have about the strategies chosen by the other players from the sets $R_j^0 = A_j$ $(j \neq i)$. Next, R_i^2 consists of all i's possible best responses given the various beliefs about the strategies chosen by other players from R_j^1 $(j \neq i)$, and so on. Evidently $\emptyset \neq R_i^k \subset R_i^{k-1}$ $(k = 1, 2, \ldots)$, as can readily be proved by induction on k. It follows that each player $i \in N$ has a well defined limit set

$$R_i := \lim_{k \to \infty} R_i^k = \bigcap_{k=0}^{\infty} R_i^k.$$

This completes the construction. Indeed, because each set A_i is finite, the limit R_i is reached after a finite number of iterations, and is non-empty. Since there is a finite number of players, moreover, there is a finite integer n, independent of i, for which $k \geq n$ implies $R_i = R_i^k$ (all $i \in N$). But then

$$R_i = B_i \left(\prod\nolimits_{j \in N \setminus \{i\}} \Delta(R_j) \right)$$

so that rationalizable strategies are best responses to possible beliefs about other players' rationalizable strategies. The following result is also easy to prove:

THEOREM. *Every strategy which is used with positive probability in some Nash equilibrium must be rationalizable. That is, if $E \subset \prod_{i \in N} \Delta(A_i)$ denotes the set of possible Nash equilibrium beliefs that players can hold about each other, and if for any player $i \in N$, the set E_i is defined as $\{ a_i \in A_i \mid \exists \pi^N = (\pi_i)_{i \in N} \in E : \pi_i(a_i) > 0 \}$, then $E_i \subset R_i$.*

PROOF: Let $\bar{\pi}^N = \prod_{i \in N} \bar{\pi}_i$ be any Nash equilibrium, where $\bar{\pi}_i \in \Delta(\bar{A}_i)$ (all $i \in N$). Let $\bar{A}_i := \{ a_i \in A_i \mid \bar{\pi}_i(a_i) > 0 \}$ be the set of strategies used by i with positive probability in this Nash equilibrium. Then $\bar{A}_i \subset B_i(\bar{\pi}_{-i})$, where $\bar{\pi}_{-i} := \prod_{j \in N \setminus \{i\}} \bar{\pi}_j$. Since $R_i^0 = A_i$, obviously $\bar{A}_i \subset R_i^0$ (all $i \in N$). Now suppose that the induction hypothesis $\bar{A}_i \subset R_i^{k-1}$ (all $i \in N$) is satisfied. Then

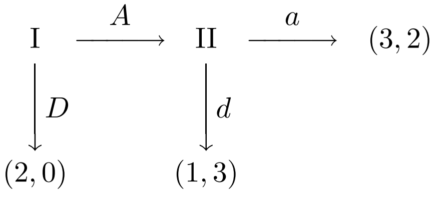
$$\bar{A}_i \subset B_i(\bar{\pi}_{-i}) \subset B_i \left(\prod_{j \in N \setminus \{i\}} \Delta(\bar{A}_j) \right) \subset B_i \left(\prod_{j \in N \setminus \{i\}} \Delta(R_j^{k-1}) \right) = R_i^k.$$

Thus $\bar{A}_i \subset R_i^k$ for $k = 1, 2, \ldots$, by induction on k, and so $\bar{A}_i \subset R_i$ as required. ∎

So all Nash strategies are rationalizable. But, as Bernheim's (1984, pp. 1024–5) Cournot oligopoly example clearly shows, other non-Nash strategies are often rationalizable too.

Since a strictly dominated strategy is never a best response, for $k = 1, 2, \ldots$ it must be true that all strictly dominated strategies for player i are removed from R_i^{k-1} in reaching R_i^k. It follows easily by induction on k that R_i^k is a subset of S_i^k, the set of i's strategies that survive k iterations of removing strictly dominated strategies. Taking the limit as $k \to \infty$ implies that R_i must be a subset of S_i^∞, the set of all i's strategies that survive iterative removal of strictly dominated strategies. Generally, however, $R_i \neq S_i^\infty$, though the two sets are always equal in two-person games (see Fudenberg and Tirole, 1991, pp. 50–52 and 63).

2.3. Refinements



A Simple Extensive Form

Rationalizability on its own fails to exclude some very implausible strategy choices. In the example shown, d is evidently an optimal strategy for player II whenever the subgame is reached. But D is player I's best response to d, and a is one of player II's best responses to D, while A is player I's best response to a. Thus all strategies are possible best responses and so rationalizable in this game. In the normal form, moreover, there are no strictly dominated strategies. Yet a is not a credible choice by Player II. In fact, a is strictly dominated in the subgame which starts with Player II's move. If only d is really rationalizable for II, however, then only D is for player I, so the only really rationalizable choices are (D, d). This illustrates *subgame rationalizability* (Bernheim, 1984, p. 1022).

The argument of the above paragraph exploited the extensive form structure of the game. But there are similar normal form arguments such as that used by (Bernheim, 1984, p. 1022) to discuss "perfect rationalizability". This paper, however, will instead make use of Pearce's (1984) idea of refining rationalizable to "cautiously rationalizable" strategies. To explain his construction, first let $\Delta^0(T)$ denote, for any finite set T, the interior of $\Delta(T)$ — i.e., the set of all probability distributions attaching positive probability to every member of T. Also, given any subset $A'_i \subset A_i$, player i's *constrained best response correspondence* is defined by

$$B_i(\pi_i \mid A'_i) := \arg\max_{a_i \in A'_i} \{\, U_i(a_i, \pi_i) \,\}$$

Now start with $\hat{A}^0_i = A_i$, the set of all i's possible strategies, and $\hat{R}^0_i := R_i$, the set of all i's rationalizable strategies. Then construct the sequences of sets $\hat{P}^k_i, \hat{A}^k_i, \hat{R}^k_i$ $(k = 1, 2, \ldots)$ recursively so that

$$\hat{P}^k_i := \prod\nolimits_{j \in N \setminus \{i\}} \Delta^0(\hat{R}^{k-1}_j),$$

while $\hat{R}^k_i$ is the set of rationalizable strategies in the restricted normal form game where each player is $i \in N$ is only allowed to choose some strategy a_i in the set

$$\hat{A}^k_i := B_i\left(\hat{P}^k_i \mid \hat{R}^{k-1}_i\right) \subset \hat{R}^{k-1}_i$$

of what Pearce called "cautious responses" within $\hat{R}^{k-1}_i$ to expectations concerning other players' strategies in the sets $\hat{R}^{k-1}_j$ $(j \neq i)$. Provided that the sets $\hat{R}^{k-1}_i$ $(i \in N)$ are non-empty, the construction will continue to yield non-empty sets $\hat{A}^k_i, \hat{R}^k_i$ $(i \in N)$ which, because of the definition of rationalizable strategies, must satisfy

$$\hat{R}_i^k = B_i\left(\bar{P}_i^k \mid \hat{A}_i^k\right) \subset \hat{A}_i^k \quad \text{where} \quad \bar{P}_i^k := \prod\nolimits_{j \in N \setminus \{i\}} \Delta(\hat{R}_j^{k-1}) = \text{cl } \hat{P}_i^k.$$

It follows by induction on k that

$$\emptyset \neq \hat{R}_i^k \subset \hat{A}_i^k \subset \hat{R}_i^{k-1} \subset \hat{A}_i^{k-1} \subset \ldots \subset \hat{R}_i^0 \subset \hat{A}_i^0 = A_i \quad (k = 1, 2, \ldots).$$

Now, repeating the argument used for rationalizable strategies shows that there must be a finite integer n for which $k \geq n$ implies that $\hat{A}_i^k = \hat{A}_i^n = \hat{R}_i^k = \hat{R}_i^n$ (all $i \in N$). The set

$$\hat{R}_i := \bigcap_{k=0}^{\infty} \hat{R}_i^k = \hat{R}_i^n \subset \hat{R}_i^0 = R_i$$

must therefore be a well defined and non-empty subset of R_i. Call $\hat{R}_i$ the set of *cautiously rationalizable strategies* for player i. Evidently $\hat{A}_i^k = \hat{R}_i$ for all $k \geq n$, and so

$$\hat{R}_i = B_i(\hat{P}_i \mid \hat{R}_i) \text{ where } \hat{P}_i := \prod\nolimits_{j \in N \setminus \{i\}} \Delta^0(\hat{R}_j) \quad (\text{all } i \in N).$$

Thus any player's cautiously rationalizable strategies are indeed (cautious) best responses to non-degenerate probability beliefs about other players' cautiously rationalizable strategies. But it is not required that all such best responses be included in the set of cautiously rationalizable strategies.

Because only strategies that are best responses to *interior* probability distributions are retained in passing from the sets $\hat{R}_i^{k-1}$ to $\hat{A}_i^k$, it follows that all *weakly* dominated strategies in $\hat{R}_i^{k-1}$ are eliminated. Thus $\hat{A}_i^k$ must be a subset of $W_i(\hat{R}_i^{k-1})$, and the product set $\hat{A}^k := \prod_{i \in N} \hat{A}_i^k$ must be a subset of $W(\hat{R}^{k-1})$, where $\hat{R}^{k-1} := \prod_{i \in N} \hat{R}_i^{k-1}$. Thereafter, all strategies among $W(\hat{R}^{k-1})$ that would be removed by iterative strict dominance must also be removed from $\hat{A}^k$ in order to arrive at strategies in the product set $\hat{R}^k$ that are rationalizable in the game where each player is $i \in N$ is artificially restricted to choosing some strategy in $\hat{A}_i^k$. It follows that $\hat{R}^k \subset S^\infty(\hat{A}^k)$. So $\hat{R}^k$ must be a subset of the set $S^\infty(W(\hat{R}^{k-1}))$ of those strategies that survive the removal of *all* weakly dominated strategies at the first stage, followed by iterated removal of all strictly dominated strategies at each later stage. Ultimately, moreover, each $\hat{R}_i$ must be a subset of the set $(W\, S^\infty)_i^\infty$ of i's strategies that survive cautious iterative removal of dominated strategies.[3]

[3] Dekel and Fudenberg (1990) justify the strategy set $S^\infty W(A^N)$ as the implication of modifying iterative deletion of *all* weakly dominated strategies at each round so as to allow each player to be

Note that, in the above construction of the sets $\hat{R}_i^k$, it is *not* enough to replace $\hat{A}_i^k$ by the whole unrestricted best response set $A_i^k := B_i(\hat{P}_i^{k-1})$. This is because, even though $\hat{R}_j^k \subset \hat{R}_j^{k-1}$ it is not generally true that $\Delta^0(\hat{R}_j^k) \subset \Delta^0(\hat{R}_j^{k-1})$, and so this alternative construction does not guarantee that $A_i^{k+1} \subset A_i^k$. Indeed, for the counterexample normal form game shown, this it would result in a different sequence of sets $\tilde{R}_i^k$ given by

$$\tilde{R}_I^k = A_I^k = \begin{cases} \{\, U, D \,\} & \text{if } k \text{ is even;} \\ \{U\} & \text{if } k \text{ is odd;} \end{cases} \qquad \tilde{R}_{II}^k = A_{II}^k = \begin{cases} \{\, L, R \,\} & \text{if } k \text{ is even;} \\ \{L\} & \text{if } k \text{ is odd.} \end{cases}$$

The difference is that, in the k-th step of Pearce's construction of the sets $\hat{R}_i^k$, only cautious responses among the sets $\hat{A}_i^k$ of surviving strategies are considered. Strategies which have already been eliminated at a previous step as not cautiously rationalizable should never be re-admitted into the set of cautiously rationalizable strategies.[4]

		II	
		L	R
I	U	1 1	1 1
	D	1 1	0 0

A Counterexample

For similar reasons, the limit sets do not generally satisfy the condition $\hat{R}_i = B_i(\hat{P}_i)$ that was investigated by Börgers and Samuelson (1992, p. 20). In particular, whereas Pearce's definition, which is the one used here, ensures both existence and uniqueness of the set of cautiously rationalizable strategies, neither is guaranteed with this alternative condition.

a little uncertain about other players' payoffs. Though the motivation would have to be different, $(W\,S^\infty)^\infty W(A^N)$ is a refinement of their concept, and is obviously related to a slightly different version of Pearce's notion of "cautious rationalizability."

[4] For this example I am indebted to Pierpaolo Battigalli. The games in examples 3 and 4 of Börgers and Samuelson (1992) could both be used to make the same point.

Nevertheless, in the appendix it will be shown that $B_i(\pi_i \mid \hat{A}_i^k) = B_i(\pi_i) \cap \hat{A}_i^k$ for all $\pi_i \in \bar{P}_i^k$. From this it follows that

$$\begin{aligned} \hat{A}_i^{k+1} &= B_i(\hat{P}_i^k \mid \hat{R}_i^k) = B_i(\hat{P}_i^k \mid B_i(\bar{P}_i^k \mid \hat{A}_i^k)\,) = B_i(\hat{P}_i^k \mid \hat{A}_i^k) \\ &= B_i(\hat{P}_i^k) \cap \hat{A}_i^k = \bigcap_{q=0}^{k} B_i(\hat{P}_i^q) \quad (k = 0, 1, 2, \ldots) \end{aligned}$$

where the last equality follows from recursion. In addition

$$\hat{R}_i^{k+1} = B_i(\bar{P}_i^k \mid \hat{A}_i^k) = B_i(\bar{P}_i^k) \cap \hat{A}_i^k = B_i(\bar{P}_i^k) \cap \left[\bigcap_{q=0}^{k-1} B_i(\hat{P}_i^q)\right] \quad (k = 0, 1, 2, \ldots).$$

Thus each set $\hat{R}_i^{k+1}$ and $\hat{A}_i^{k+1}$ consists of *unconstrained* best responses, but only those which remain “cautiously rationalizable” after k steps of the construction.

2.4. Rationalizable Expectations

In the following, I shall use the obvious notation

$$R_{-i} := \prod\nolimits_{j \in N \setminus \{i\}} R_j; \quad \hat{R}_{-i} := \prod\nolimits_{j \in N \setminus \{i\}} \hat{R}_j$$

for the sets of all possible profiles of rationalizable (resp. cautiously rationalizable) strategies that the players other than i are able to choose. Now, if a strategy $s_i \in R_i$ is rationalizable, it is because there exists a probability distribution $\pi_i \in \Delta(R_{-i})$ over other players' rationalizable strategies such that $s_i \in B_i(\pi_i)$. Similarly, if a strategy $s_i \in \hat{R}_i$ is cautiously rationalizable, it is because there exists an interior probability distribution $\pi_i \in \Delta^0(\hat{R}_{-i})$ over other players' cautiously rationalizable strategies such that $s_i \in B_i(\pi_i)$. If player i has a subjective probability distribution satisfying $\pi_i \in \Delta(R_{-i})$, therefore, then i will be described as having *rationalizable expectations.* This accords with the terminology introduced by Bernheim (1984, p. 1025). Similarly, if in fact $\pi_i \in \Delta^0(\hat{R}_{-i})$, then i will be described as having *cautiously rationalizable expectations.* Moreover, given any $\pi_i \in \Delta(R_{-i})$ for which the strategy $s_i \in B_i(\pi_i)$ is a best response, the expectations π_i will be described as *rationalizing* that strategy.

3. Correlated Strategies

3.1. Background

Aumann (1987) has proposed an interesting extension to Nash equilibrium. A *correlated equilibrium* is a joint probability distribution $\bar{\pi} \in \Delta(A^N)$, possibly with correlation between different players' strategies, with the property that, for all $i \in N$, whenever $\bar{a}^N \in A^N$ satisfies $\bar{\pi}(\bar{a}^N) > 0$, then

$$\bar{a}_i \in \arg\max_{a_i \in A_i} \{\, \mathbb{E}_{\pi(\cdot|\bar{a}_i)} v_i(a_i, a_{-i}) \,\}.$$

Here $\pi(\cdot|\bar{a}_i) \in \Delta(A_{-i})$ denotes the conditional distribution of other players' strategies given that player i chooses $\bar{a}_i$. As Aumann points out, this notion of correlated equilibrium is consistent with the two players in Battle of the Sexes choosing (A, a) with probability half, and (B, b) with probability half.

Though Aumann argues quite persuasively for this "coarsening" of the usual Nash equilibrium concept, it does create some serious conceptual problems. For how can the two players ensure that their strategy choices are perfectly correlated, as they must be in the above correlated equilibrium of Battle of the Sexes? Effectively, after all, player R is required to believe that whatever causes her to choose A also causes C to choose a, and *vice versa*. Similarly for B and b. These beliefs appear highly implausible in the absence of a correlation device. Of course, if there really is a correlation device, such as traffic lights at a busy cross-roads, then that should be modelled as an explicit feature of the game, so that players' strategies can be conditioned by the information provided by the correlation device. Brandenburger and Dekel (1987) do suggest that it is legitimate to consider correlated strategies even where it is understood only implicitly that there may be a correlation device which is not explicitly modelled within the game. This, however, seems unsatisfactory because it is not clear how much correlation is really possible on the basis of the unmodelled device.

Actually, the beliefs behind correlated equilibria are somewhat reminiscent of those that underlie "causal decision theory", as discussed by Nozick (1969) and others — see Campbell and Sowden (1985) as well as Gärdenfors and Sahlin (1988, Part V). There is, however, an important difference between correlated equilibrium and causal decision theory. In Newcomb's problem, for instance, it is assumed that one player's choice can somehow *cause* the choice of a second player to change, even though the second player has to move before observing the first player's choice. If such causation were really possible, it would obviously be rational for players to take it into account. Therefore, causal decision theory suggests that each player $i \in N$ should choose an $\bar{a}_i$ to maximize $\mathbb{E}_{\pi(\cdot|a_i)} v_i(a_i, a_{-i})$ with respect to a_i, instead of maximizing $\mathbb{E}_{\pi(\cdot|\bar{a}_i)} v_i(a_i, a_{-i})$ as the above definition of correlated equilibrium would require. The difference is whether the relevant expectations are described by $\pi(\cdot|a_i)$, which the action a_i causes to vary, or by $\pi(\cdot|\bar{a}_i)$ instead, which remains fixed even as a_i varies within the set A_i.

3.2. Correlated Rationalizability

$$
\begin{array}{cc}
 & \text{II} \\
\text{I} &
\begin{array}{|c|c|c|}
\hline
 & l & r \\
\hline
U & \begin{array}{lll}1 & & \\ & 1 & \\ & & 1\end{array} & \begin{array}{lll}0 & & \\ & 0 & \\ & & 0\end{array} \\
\hline
D & \begin{array}{lll}0 & & \\ & 0 & \\ & & 0\end{array} & \begin{array}{lll}1 & & \\ & 1 & \\ & & 1\end{array} \\
\hline
\end{array}
\end{array}
\xleftarrow{\alpha} \text{III} \xrightarrow{\beta}
\begin{array}{cc}
 & \text{II} \\
\text{I} &
\begin{array}{|c|c|c|}
\hline
 & l & r \\
\hline
U & \begin{array}{lll}1 & & \\ & 1 & \\ & & 0\end{array} & \begin{array}{lll}0 & & \\ & 0 & \\ & & 1\end{array} \\
\hline
D & \begin{array}{lll}0 & & \\ & 0 & \\ & & 1\end{array} & \begin{array}{lll}1 & & \\ & 1 & \\ & & 0\end{array} \\
\hline
\end{array}
\end{array}
$$

Correlated Rationalizability

Neither causal decision theory nor correlated equilibrium makes much sense to me, at least. Still, as Pearce (1984, p. 1048), Bernheim (1986) and many others have realized, that does not rule out the possibility of each player's strategy choice being rationalized by correlated beliefs about all the *other* players' strategies. Indeed, consider the game shown, in which player III's intial choice of strategy α or β determines whether the left or the right tri-matrix will occur. In this game, it makes very good sense for player III to take a view regarding how likely it is that players I and II can co-ordinate their strategies and choose either (U, l) or (D, r) with high probability. A player III who thinks this is likely will want to choose α and go to the left hand tri-matrix; but one who thinks it unlikely will want to choose β and go to the right hand tri-matrix instead. Constraining III's beliefs to exclude any such possibility of correlation does not seem reasonable, since player III may believe that there really is some common causation behind I and II's behaviour, even if there is no correlation device or other means of communication. For instance, I and II may be identical twins, who are always observed to make matching choices. Or player III may know that I and II were able to communicate at some time in the past, before the start of this game.[5]

The important point to realize is this. With common expectations, if any two players $i, j \in N$ have correlated strategy choices, then i must believe that the likelihood of j's choice is conditioned on i's own choice. Without common expectations, however, any two players can believe that what they choose is independent of the other's strategy, even though any third player may regard their choices as stochastically dependent. As an example, identical twins probably feel that they are always choosing independently, even though outside observers and third players only see them making identical choices. And suppose there is a group of people who, because they have agreed in the past about what to do in a game which confronts them later, are believed by outsiders to be playing correlated strategies. Nevertheless, this group will not actually be able to correlate their choices in that game, unless its structure explicitly allows further communication.

Allowing correlated beliefs about the strategies of others makes more strategies rationalizable, of course. Indeed, it weakens rationalizability so that it becomes what Brandenbuger and Dekel (1987) called *correlated rationalizability*. It has become well known that this is equivalent to iterated removal of strictly dominated strategies — see, for example, Fudenberg and Tirole (1991, p. 52). In other words, the set R_i of rationalizable strategies expands to become equal to S_i^∞. Allowing correlated beliefs also weakens cautious rationalizability so that $\hat{R}_i$ becomes equal to $(W\,S^\infty)_i^\infty$ — as can be seen from the fact (Pearce, 1984, p. 1049, Lemma 4) that a strategy is not weakly dominated iff it is a cautious best response. Pearce proves this only for two-person games, but remarks how the same proof would work for n-person games if correlated beliefs about other players' strategies were allowed.

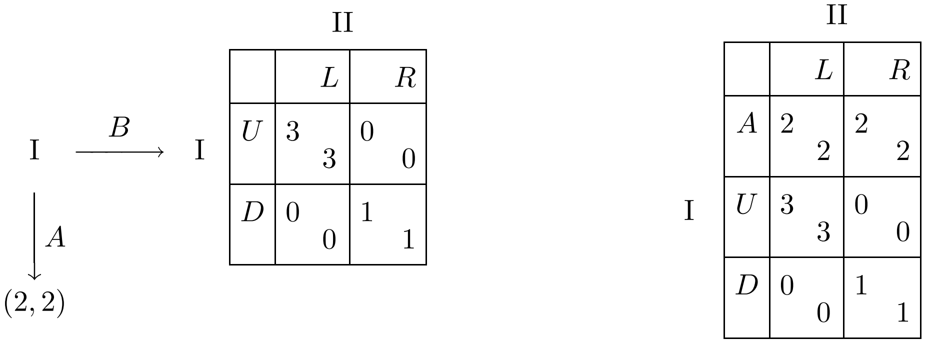
To summarize, therefore, we have the following characterization:

[5] In fact, as Pierpaolo Battigalli has pointed out to me, in this game where each player has only two strategies, the sets of rationalizable and correlated rationalizable strategies must be identical. So must the the sets of cautiously rationalizable and correlated cautiously rationalizable strategies. The reason is that, whenever a player has only two strategies, one of those strategies is never a best response iff it is strictly dominated by the other (and is never a cautious best response iff it is weakly dominated by the other). So the process of successively eliminating never (cautious) best responses is equivalent to that of iterative deletion of dominated strategies. In view of the theorem stated below, this establishes the claim. Yet the discussion in the main text concerns only the expectations regarding other players' strategies that may rationalize a particular rationalizable strategy. The set of such expectations is obviously expanded by allowing correlation.

THEOREM. *A strategy is correlated rationalizable if and only if it survives iterative removal of strictly dominated strategies, and is cautiously correlated rationalizable if and only if it survives cautious iterative removal of dominated strategies.*

4. Forward Induction and Conditional Rationalizability

Forward induction was first introduced by Kohlberg and Mertens (1986). The discussion here begins, therefore, with a special case of the example they present in Figure 3 on p. 1008. I have taken $x = 0$ to make the example into a "team decision problem" (Marschak and Radner, 1972) with both players sharing the same payoffs. An earlier example by Kohlberg with some of the same features (though not a team decision problem) was discussed by Kreps and Wilson (1982, p. 885), Bernheim (1984, p. 1023), Pearce (1984, p. 1044) and McClennan (1985).



A Team Version of Kohlberg's Example

In this example, the obvious strategy choices (U, L) form one Nash equilibrium, but (A, R) is another. Indeed, both (U, L) and (A, R) are sequential equilibria. The strategies A, U for I and L, R for II are therefore rationalizable. Now, for player I strategy D is strictly dominated by A. Yet, after the strictly dominated strategy D has been eliminated, R becomes a weakly dominated strategy for player II, and so an incautious best response. After it has been eliminated next, then A becomes a strictly dominated strategy for player I. So iterative deletion of all weakly dominated strategies among those that remain rationalizable leads ultimately to (U, L) as the only possible outcome. By our previous theorem, this is also the only cautiously rationalizable outcome.

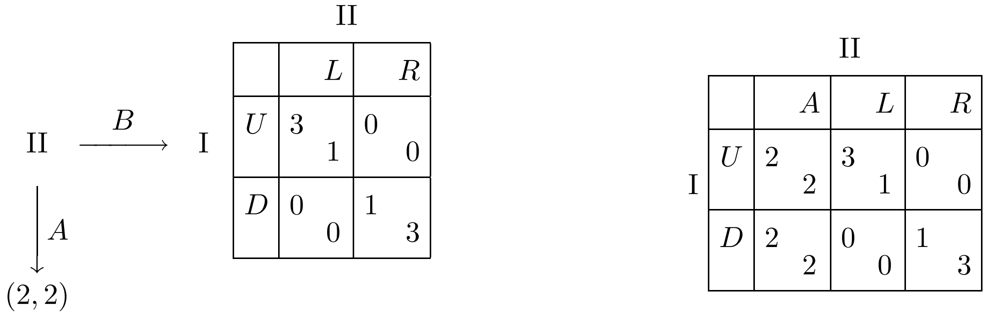
The forward induction argument of Kohlberg and Mertens (1986, p. 1013), however, seems quite different, at least to begin with. They write that "a subgame should not be treated as a separate game, because it was preceded by a very specific form of preplay communication — the play leading to the subgame." Yet, according to the standard view of Nash equilibrium and, more particularly, of sequential equilibrium, there is no reason for such preplay communication to be relevant at all. Indeed, the sequential equilibrium which the players are supposed to be following prescribes beliefs in every possible subgame, and as long as these fulfil all the requirements of sequential equilibrium, there is nothing more that equilibrium theory can say.[6] In fact, Kohlberg and Mertens expect player II, when required to move, to be influenced by the fact that player I has given up an opportunity to get a payoff of 2. Though their argument is rather sketchy, it seems that we are expected to conclude that it is reasonable for player II, faced with the move, to believe that player I was expecting to get at least 2 in the subgame. Yet, according to sequential equilibrium theory, there really is no reason for player II to abandon the beliefs that sustain (A, R) as the anticipated sequential equilibrium; instead, I's playing B is seen as a mere "tremble".

It seems to me, therefore, that this kind of forward induction argument makes much more sense when we think of rationalizable strategies. The same point, in fact, was suggested by Bernheim and Pearce themselves, and also appears more recently in Battigalli (1991). The theory of rationalizability allows us to recognize that player II may not be certain after all what will happen in the subgame where there is a move to make, and will be looking for clues concerning what is likely to happen. In the subgame all strategies are rationalizable, and so II has no information within the subgame itself which helps decide whether I is more likely to play U or D. However, there is crucial information from outside the subgame — namely, the fact that I has chosen B rather than A, so giving up the opportunity to get 2 for sure. In the framework of rationalizability, the above quotation from Kohlberg and Mertens makes excellent sense — much more, it seems to me, than if one considers only sequential equilibria, where it is presumed that players already have equilibrium beliefs, and so cannot possibly be influenced by any form of "preplay communication." In fact, knowing

[6] It should be emphasized that neither Kreps and Wilson (1982) nor McClennan (1985) really contradict this, since they use the earlier version of Kohlberg's example only to argue that it may be desirable to impose additional restrictions on players' sequential equilibrium beliefs.

that player I has foregone the opportunity to get a payoff of 2, it is plausible for player II to believe that player I is expecting a payoff of at least 2 in the subgame. This is only possible, however, if player I expects to play U in the subgame, regarding it as sufficiently likely that player II will play L. So player II, also expecting I to play U in the subgame, should play L.

The conclusion is that, once the subgame has been reached, only (U, L) is "conditionally rationalizable," in the obvious sense that it is rationalized by rationalizable expectations satisfying the condition that player I's expected payoff in the subgame should be at least 2. Of course, if player I foresees that this will be the outcome of the subgame, then player I will indeed choose to play B and then U, and so the "sensible" outcome is indeed achieved.



	L	R
U	3 1	0 0
D	0 0	1 3

	A	L	R
U	2 2	3 1	0 0
D	2 2	0 0	1 3

Battle of the Sexes with a Correlation Option

Another rather similar example is a slight variation on one that was first discussed by van Damme (1989, p. 479).[7] The game can be regarded as an extension of Battle of the Sexes in which player II is given the option of setting up beforehand a correlation device that will ensure the symmetric correlated equilibrium payoffs $(2, 2)$. For this example, the set of cautiously rationalizable strategies can be found by iteratively deleting dominated strategies as follows. First, A strictly dominates L for player II. But this is the only strict dominance relation, so U, D remain as rationalizable strategies for player I, as do A, R for player II. However, after L has been eliminated, D weakly dominates U for player I; again, this is the only dominance relation at this stage. Finally, after U has also been eliminated,

[7] The only difference is that van Damme lets player II have a payoff of 5 instead of 2 if player I chooses the strategy A. It is easy to see that this makes no difference to the argument presented below for this particular game.

R strictly dominates A for player II, and so we are left with (D, R) as the only possible outcome.

In this example, the forward induction argument leads to the same result. If player II foregoes the correlation option and chooses B, this can only be rationalized by the belief that II expects to get a payoff of at least 2 in the subgame. Thus player II must be intending to play R in the belief that, with probability at least $\frac{2}{3}$, player I will choose D. This leads to (D, R) as the only possible conditionally rationalizable outcome in the subgame. So player II is induced to enter the subgame and play Battle of the Sexes without a correlation device in the expectation of achieving this preferred outcome.

The same game can be used to illustrate the inconsistency of this kind of forward induction or iterated elimination of dominated strategies with Harsanyi and Selten's theory of equilibrium selection[8] — see Harsanyi and Selten (1988, Section 10.8) and van Damme (1990) for other examples making a similar point. For their theory imposes symmetry upon the (unique) solution to all symmetric games such as Battle of the Sexes. It follows that the only Harsanyi-Selten equilibrium in the subgame is the mixed strategy symmetric equilibrium $(\frac{3}{4}U + \frac{1}{4}D, \frac{1}{4}L + \frac{3}{4}R)$ yielding expected payoffs $(\frac{3}{4}, \frac{3}{4})$. In the full game with a correlation option this would encourage player I to choose A, and so lead to the (mixed strategy) equilibrium $(A, \frac{1}{4}L + \frac{3}{4}R)$. This contradicts the conclusion of the previous paragraph, of course, and so demonstrates the claimed inconsistency.

The game we have just been discussing is actually a subgame of another, considered by Dekel and Fudenberg (1990, Fig. 7.1, pp. 265–6), which can be regarded as a version of Battle of the Sexes allowing two correlation options. Indeed, compared with the previous example, player I has now also been given the option of using a correlation device, but a somewhat costly one, since using it reduces the payoffs to $(1.5, 1.5)$. As is easy to check, cautious iterated deletion of dominated strategies now leads to (X, R) as the unique cautiously rationalizable outcome. There is one other Nash equilibrium (U, A) that also happens to be subgame perfect.

Yet Dekel and Fudenberg claim that (D, A) is also reasonable in this example. The reason they give is that, "if player II accepts the [(X, R)] solution (which is based on the

[8] Such inconsistencies were kindly pointed out to me by Giacomo Costa, but with a different example similar to the one used by Harsanyi and Selten themselves.

$$\begin{tikzcd}
\mathrm{I} \arrow[r, "Y"] \arrow[d, "X"] & \mathrm{II} \arrow[r, "B"] \arrow[d, "A"] & \mathrm{I} \\
(1.5, 1.5) & (2, 2) &
\end{tikzcd}$$

	II	
	L	*R*
U	3 1	0 0
D	0 0	1 3

		II		
		A	*L*	*R*
I	*X*	1.5 1.5	1.5 1.5	1.5 1.5
	U	2 2	3 1	0 0
	D	2 2	0 0	1 3

Battle of the Sexes with Two Correlation Options

intuition of forwards induction) and then is given the opportunity to play, II must conclude that 'something basic has changed,' and II might conclude that I's payoffs will lead I to violate the [(D, R)] outcome in the subgame." In fact, there is a really interesting tension here between I's attempt, by playing Y, to convince II that player I expects a payoff of at least 1.5, and II's attempt, by playing B, to convince player I that II expects a payoff of at least 2. Without a correlation device, there is simply no way to fulfill both expectations simultaneously. Quite a likely outcome of the second subgame is that player I will choose U while II plays R, each thinking it likely that the other will play the alternative strategy, and each expecting a higher payoff than what would result from their foregone opportunities to play X and A respectively.

Nevertheless, the outcome in which I plays X still makes the most sense to me. After choosing Y first, will player I really proceed with U if II plays B? Or is player I more likely to realize at the last minute that II really must be intending R, and so play D in the end? It seems that II has an advantage in the first subgame from having the last move before the second subgame starts.[9] Anyway, compared to standard equilibrium theory,

[9] Ben-Porath and Dekel (1992, p. 46) have an example which displays a similar "vulnerability to counter-signals."

II

		L_B	R_B
I	U	3 1	-2 0
	D	-2 0	-1 5

$\xleftarrow{B}$ I $\xrightarrow{N}$

II

		L_N	R_N
I	U	5 1	0 0
	D	0 0	1 5

II

		L_BL_N	L_BR_N	R_BL_N	R_BR_N
I	BU	3 1	3 1	-2 0	-2 0
	BD	-2 0	-2 0	-1 5	-1 5
	NU	5 1	0 0	5 1	0 0
	ND	0 0	1 5	0 0	1 5

Ben-Porath and Dekel's "Money Burning" Example

rationalizability clearly allows a much richer discussion of what players can reasonably expect of each other.

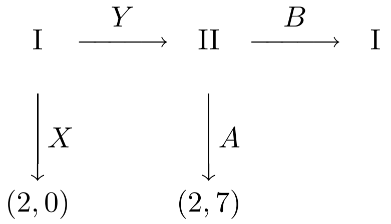
In the examples discussed so far, it was easy to find a forward induction argument yielding the same unique outcome as iterated elimination of all weakly dominated strategies. This will not be so easy in the next example. In the form given here, this is originally due to Ben-Porath and Dekel (1992), though it is based on an idea due to van Damme (1989, pp. 488–90) that is also discussed by Myerson (1991, pp. 194–5); see also Fudenberg and Tirole (1991, pp. 461–4).[10] The game begins with player I having a choice between B, to be interpreted as "burning money" which leads to a loss of 2 units of utility for player I, and N, to be interpreted as not burning money. After N the basic payoffs are given by the right-hand bi-matrix, which is a modified form of Battle of the Sexes. After B the payoffs are given by the left-hand bi-matrix. Compared with the right-hand bi-matrix, I's payoffs have all been reduced by 2, but II's payoffs are exactly the same.

[10] Actually, one of the payoffs given in Ben-Porath and Dekel (1992, Fig. 1.2) is a 4, when a 5 is clearly intended. The change is irrelevant to the argument, however.

In the corresponding normal form of this game, BU, BD, NU, ND are the four strategies for player I, while $L_B L_N, L_B R_N, R_B L_N, R_B R_N$ are the four strategies for player II. Of I's strategies, ND strictly dominates BD and so BD can be eliminated, but all other strategies of both players remain rationalizable. After BD has been eliminated, $L_B L_N$ weakly dominates $R_B L_N$ for player II, and $L_B R_N$ weakly dominates $R_B R_N$. So $R_B L_N$ and $R_B R_N$ are eliminated, leaving $L_B L_N$ and $L_B R_N$. Once it is known that player II will choose L_B, however, BU strictly dominates ND for player I and so ND can be eliminated. This leaves only BU and NU as possible cautiously rationalizable strategies for player I. However, of player II's remaining strategies, $L_B L_N$ weakly dominates $L_B R_N$, leaving $L_B L_N$ as player II's only cautiously rationalizable strategy. This leaves NU as player I's only cautiously rationalizable strategy, and $(NU, L_B L_N)$ as the only cautiously rationalizable outcome, with $(9, 6)$ as the resulting payoffs.

To establish $(NU, L_B L_N)$ as the only conditionally rationalizable outcome in this game, the corresponding forward induction argument has to pass backwards and forwards between the two different possible subgames which can occur after I's first move. Notice first how, by choosing ND, player I can guarantee a payoff of at least 0. So if player I were to play B, it could only be in the expectation of obtaining at least 0 in the left-hand subgame. With conditionally rationalizable strategies, this is only possible if player I were to choose U expecting II to play L_B. Thus, if player I were to choose B, it would be in the expectation of getting a payoff of 3. Now, however, if player I were to choose N instead, this could only be rationalized if player I were expecting a payoff of at least 3 in the right hand subgame following that first move. This, however, is only possible if I intends to play U and expects II to choose L_N with a sufficiently high probability. Conditional rationalizability then requires player II to choose L_N, so yielding $(NU, L_B L_N)$ as the only possible outcome of forward induction.

It should be pointed out that Ben-Porath and Dekel regard it as surprising that giving player I the opportunity to "burn money" in this way should confer an advantage in the subgame that occurs after no money has been burnt. They also consider a large class of games in which a similar phenomenon arises. But is this really any more surprising than the possibility which Schelling (1960, p. 24) noticed — namely, that the potential buyers of a house could influence the outcome of a bargaining process with the sellers by first making a large enough bet with an outside party that they would not pay more than a specified price? Of course, the (foregone) opportunity to burn money does only affect conditionally rationalizable beliefs, whereas Schelling's house buyers' bet affects payoffs directly. Yet both affect *expected* payoffs, which are what rationalize behaviour.



II

	L	R
U	9 0	0 1
D	1 0	1 8

II

		A	BL	BR
	X	2 0	2 0	2 0
I	YU	2 7	9 0	0 1
	YD	2 7	1 0	1 8

Myerson's Example

My last example is due to Myerson (1991, pp. 192–3). It illustrates that there is as yet no generally accepted concept of forward induction. To quote (with some changes to reflect different notation):

> *Unfortunately, some natural forward-induction arguments may be incompatible with other natural backward-induction arguments. [In the example shown, a] forward-induction argument might suggest that, because YD is a weakly dominated strategy for player I, player II would expect player I to choose U [in the last subgame] if he chose Y [originally]; so player II should choose A. On the other hand, backward induction determines a unique subgame-perfect equilibrium (XD, BR) in which player II would choose B [for his first move]. ... Forward induction corresponds to interatively eliminating weakly dominated strategies in the order: first YD for player I, then BL and BR for player II, leaving the equilibria (X, A) and (YU, A).*

In fact (X, A) is the unique cautiously rationalizable outcome. For in the game as a whole there are no strictly dominated strategies, so all strategies are rationalizable. Thereafter, on the first round of eliminating weakly dominated strategies, both YD for player I and BL for player II get eliminated. This leaves (U, R) as the only possible outcome of the last subgame. Then there are still no strictly dominated strategies. But both YU for player I and BR for player II get eliminated on the second round of eliminating weakly dominated strategies, leaving (X, A) as the only possibility.

As for Myerson's alternative outcome (YU, A), however, it is really only credible as an outcome of this game if more reliance is placed on the logic of forward induction on the basis of player I's first move than on that of either subgame rationalizability or forward induction on the basis of player II's first move. For R strictly dominates L in the last subgame, so subgame rationalizability leaves (D, R) as the only possible outcome of the last subgame. Only if player II believes rather strongly that player I is likely to choose U in the subgame, presumably expecting II to play the dominated strategy L, will II choose A over B. Moroever, the logic of forward induction allows player I to infer from the choice of B that II's expected payoff in the subgame is at least 7, so reinforcing the claim that only (D, R) is really rationalizable in the last subgame. This leads me to claim that (XD, BR) is the only convincing outcome of this game. Of course, this outcome both differs from and makes more sense than the unique cautiously rationalizable outcome (X, A). So forward induction can suggest different strategy combinations from cautious rationalizability, even though the consequences of (XD, BR) and (X, A) are, of course, entirely the same in this example.

That concludes the series of examples which make up this section. They show how powerful forward induction arguments can be, but they still need to be formalized for general extensive and normal form games before any of their general implications can be deduced. As yet I know of no general theorems relating forward induction to cautious rationalizable. Nor do I know of an accepted definition of forward induction, even though Hillas (1990, p. 1368), for example, offers one that works well enough for his purposes.

One clear conclusion does emerge, however. In each of the subgames discussed in this section, what is conditionally rationalizable after applying forward induction depends on which options and what resulting payoffs the players in the subgame had renounced beforehand. An adequate description of the subgame therefore requires specifying these foregone options, as well as the usual extensive form of the subgame. By itself that may not be too controversial even though, as van Damme (1989) points out, it amounts to admitting that phenomena like "sunk costs" may be relevant after all in subgames. Notice, however, that a subgame is itself a game. The fact that the extensive form is an inadequate description of a subgame therefore suggests that it may not be an adequate description of a full game either. Indeed, faced with a particular game in extensive form to analyse, the examples of this section suggest that the game theorist should at least be asking what outside options the players may have given up in the past, before the game even started, since these will influence players' (conditionally) rationalizable expectations concerning what should happen in the course of the game itself. Of course, foregone options are only one of a host of extraneous features which may help to determine rationalizable expectations in the game. One of the benefits of replacing equilibrium analysis with considerations of rationalizability is precisely that such considerations, which do seem to influence the outcomes of actual games, can be brought into the game-theoretic analysis. The obvious disadvantage is that much less precise conclusions are possible, yet such precision should not replace realism.

5. Conclusions

Harsanyi and Selten (1988, pp. 342–3), amongst others, clearly point out how equilibrium theory sets itself the worthy goal of determining both what actions and also what "rational" or equilibrium expectations should arise in general non-cooperative games. Usually, this goal is not achievable, even in games with unique Nash equilibria (Bernheim, 1986). Instead, the less demanding requirement of rationalizability seems much more reasonable, since it permits each player to be uncertain about what other players believe, unlike the strait-jacket of equilibrium. Besides, it is just enough to allow individual players to have appropriate "rationalizable" expectations concerning everything about which they are uncertain in the game, including both the strategies and the expectations of all the other players (see Tan and Werlang, 1988). Correlated strategies and forward induction also appear to make more sense in the context of rationalizable rather than rational expectations. In addition, when each player is allowed to have correlated expectations regarding the strategies of all the other players, then rationalizable strategies are precisely those which survive iterated removal of strictly dominated strategies, whereas "cautiously" rationalizable strategies are those which survive "cautious" iterated removal of dominated strategies. And the intuition of forward induction seems to be well captured in a form of "conditional rationalizability" that excludes all beliefs giving a lower expected payoff to any player than the best outside option which that player is known to have passed up beforehand.

Rationalizability therefore removes some difficulties that equilibrium theory has created unnecessarily. Several anomalies remain, however. For example, the prevalence of tit-fortat in finite repetitions of prisoners' dilemma is no more consistent with rationalizability than with equilibrium theory. This and other related anomalies may well need explaining by some "bounded" version of rationalizability instead, recognizing players' inability to formulate excessively complicated models of the game they are playing (cf. Hammond, 1990). In particular, it is once again easier to contemplate bounds on rationalizable models rather than on rational or equilibrium models that represent each player's expectations concerning the game. As always, however, there is much work left to do, partly on bounded rationalizability, but also on unbounded rationalizability which was the subject of this paper.

Appendix

LEMMA. *$B_i(\pi_i \mid \hat{A}_i^k) = B_i(\pi_i) \cap \hat{A}_i^k$ for all $\pi_i \in \bar{P}_i^k$.*

PROOF: By induction on k. For $k = 0$ one has $B_i(\pi_i \mid \hat{A}_i^0) = B_i(\pi_i) \cap \hat{A}_i^0 = B_i(\pi_i)$ because $\hat{A}_i^0 = A_i$. As the induction hypothesis, suppose that $B_i(\tilde{\pi}_i \mid \hat{A}_i^{k-1}) = B_i(\tilde{\pi}_i) \cap \hat{A}_i^{k-1}$ for all $\tilde{\pi}_i \in \bar{P}_i^{k-1}$.

Suppose now that $\pi_i \in \bar{P}_i^k$ and $\hat{a}_i \in \hat{A}_i^k \setminus B_i(\pi_i)$. Then there exist $\epsilon > 0$ and $a_i \in A_i$ such that

$$U_i(a_i, \pi_i) > U_i(\hat{a}_i, \pi_i) + 2\epsilon.$$

Because $\bar{P}_i^k$ is a subset of the closure of $\hat{P}_i^{k-1}$, there exists a sequence $\langle \pi_i^\nu \rangle_{\nu=1}^\infty$ in $\hat{P}_i^{k-1}$ which converges to $\pi_i \in \bar{P}_i^k$. Construct a corresponding sequence

$$a_i^\nu \in B_i(\pi_i^\nu \mid \hat{A}_i^{k-1}) = B_i(\pi_i^\nu) \cap \hat{A}_i^{k-1} \quad (\nu = 1, 2, \ldots),$$

where the equality follows from the induction hypothesis. After choosing an appropriate subsequence if necessary, there will exist $a_i^* \in \hat{A}_i^{k-1}$ such that $a_i^\nu = a_i^*$ for all large ν, while $\pi_i^\nu \to \pi_i$ as $\nu \to \infty$.

For all large ν, since $a_i^\nu \in B_i(\pi_i^\nu)$, one has

$$U_i(a_i^*, \pi_i^\nu) = U_i(a_i^\nu, \pi_i^\nu) \geq U_i(a_i, \pi_i^\nu) > U_i(\hat{a}_i, \pi_i^\nu) + \epsilon.$$

Taking the limit as $\nu \to \infty$ implies that $U_i(a_i^*, \pi_i) \geq U_i(\hat{a}_i, \pi_i) + \epsilon$. Since

$$a_i^* = a_i^\nu \in B_i(\pi_i^\nu \mid \hat{A}_i^{k-1}) \subset B_i(\hat{P}_i^{k-1} \mid \hat{A}_i^{k-1}) = \hat{A}_i^k$$

for all large ν, it follows that $\hat{a}_i \notin B_i(\pi_i \mid \hat{A}_i^k)$.

So, for all $\pi_i \in \bar{P}_i^k$, it has been proved that $\hat{a}_i \in \hat{A}_i^k \setminus B_i(\pi_i)$ implies $\hat{a}_i \notin B_i(\pi_i \mid \hat{A}_i^k)$. Hence $B_i(\pi_i \mid \hat{A}_i^k) \subset B_i(\pi_i) \cap \hat{A}_i^k$. Since it is trivially true that $B_i(\pi_i) \cap \hat{A}_i^k \subset B_i(\pi_i \mid \hat{A}_i^k)$ even when the former set is empty, it follows that $B_i(\pi_i \mid \hat{A}_i^k) = B_i(\pi_i) \cap \hat{A}_i^k$ for all $\pi_i \in \bar{P}_i^k$. The proof by induction is complete. ▌

References

Aumann, R.J. 1987. Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica* 55:1–18.

Battigalli, P. 1991. On Rationalizability in Extensive Games. Preprint, Istituto di Economia Politica, Università Commerciale "L. Bocconi", Milan; to appear as ch. 15 of this volume.

Ben-Porath, E. and E. Dekel. 1992. Signaling Future Actions and the Potential for Sacrifice. *Journal of Economic Theory* 57:36–51.

Bernheim, B.D. 1984. Rationalizable Strategic Behavior. *Econometrica* 52:1007–1028.

Bernheim, B.D. 1986. Axiomatic Characterizations of Rational Choice in Strategic Environments. *Scandinavian Journal of Economics* 88:473–488.

Börgers, T. and L. Samuelson. 1992. "Cautious" Utility Maximization and Weak Dominance. *International Journal of Game Theory* 21:13–25.

Brandenburger, A. and E. Dekel. 1987. Rationalizability and Correlated Equilibria. *Econometrica* 55:1391–1402.

Campbell, R. and L. Sowden (eds.). 1985. *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press.

Dekel, E. and D. Fudenberg. 1990. Rational Behavior with Payoff Uncertainty. *Journal of Economic Theory* 52:243–267.

Farquharson, R. 1969. *Theory of Voting*. Oxford: Basil Blackwell.

Fudenberg, D. and J. Tirole. 1991. *Game Theory*. Cambridge, Mass.: MIT Press.

Gärdenfors, P. and N.E. Sahlin (eds.). 1988. *Decision, Probability and Utility: Selected Readings*. Cambridge: Cambridge University Press.

Hammond, P.J. 1990. A Revelation Principle for (Boundedly) Bayesian Rationalizable Strategies. European University Institute, Working Paper No. ECO 90/4.

Harsanyi, J.C. 1967–8. Games with Incomplete Information Played by 'Bayesian' Players, I–III. *Management Science* 14:159–182, 320–334, 486–502.

Harsanyi, J.C. and R. Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, Mass.: MIT Press.

Hillas, J. 1990. On the Definition of the Strategic Stability of Equilibria. *Econometrica* 58:1365–1390.

Kohlberg, E. and J.-F. Mertens. 1986. On the Strategic Stability of Equilibria. *Econometrica* 54:1003–1037.

Kreps, D. and R. Wilson. 1982. Sequential Equilibrium. *Econometrica* 50:863–894.

Luce, R.D. and H. Raiffa. 1957. *Games and Decisions: Introduction and Critical Survey*. New York: John Wiley.

Marschak, J. and R. Radner. 1972. *Economic Theory of Teams*. New Haven: Yale University Press.

McLennan, A. 1985. Justifiable Beliefs in Sequential Equilibrium. *Econometrica* 53:889–904.

Morgenstern, O. 1928. *Wirtschaftsprognose: Eine Untersuchung ihrer Voraussetzungen und Möglichkeiten* [Economic Forecasting: An Investigation of its Presuppositions and Possibilities]. Vienna: Julius Springer.

Morgenstern, O. 1935. Vollkommene Voraussicht und wirtschaftliches Gleichgewicht [Perfect Foresight and Economic Equilibrium]. *Zeitschrift für Nationalökonomie* 6:337–357.

Myerson, R.B. 1991. *Game Theory: Analysis of Conflict*. Cambridge, Mass.: Harvard University Press.

Nozick, R. 1969. Newcomb's Problem and Two Principles of Choice. In *Essays in Honor of C.G. Hempel*, edited by N. Rescher *et al.* Dordrecht: D. Reidel pp. 114–146.

Pearce, D. 1984. Rationalizable Strategic Behavior and the Problem of Perfection. *Econometrica* 52:1029–1050.

Schelling, T.C. 1960, 1980. *The Strategy of Conflict*. Cambridge, Mass.: Harvard University Press.

Tan, T.C.-C. and S.R.da C. Werlang. 1988. The Bayesian Foundations of Solution Concepts of Games. *Journal of Economic Theory* 45:370–391.

van Damme, E. 1989. Stable Equilibria and Forward Induction. *Journal of Economic Theory* 48:476–498.

van Damme, E. 1990. On Dominance Solvable Games and Equilibrium Selection Theories. CentER Discussion Paper No. 9046, University of Tilburg.

von Neumann, J. and O. Morgenstern. 1943, 1953. *Theory of Games and Economic Behavior (3rd. edn.)*. Princeton: Princeton University Press.