MATHIAS RISSE

WHAT IS RATIONAL ABOUT NASH EQUILIBRIA?

ABSTRACT. Nash Equilibrium is a central concept in game theory. It has been argued that playing Nash Equilibrium strategies is rational advice for agents involved in one-time strategic interactions captured by non-cooperative game theory. This essay discusses arguments for that position: von Neumann–Morgenstern's argument for their minimax solution, the argument from self-enforcing agreements, the argument from the absence of probabilities, the transparency-of-reasons argument, the argument from regret, and the argument from correlated equilibrium. All of these arguments either fail entirely or have a very limited scope. Whatever the use of Nash Equilibrium is, therefore, it is not useful as a rational recommendation in one-time strategic interactions. This is good news for Bayesians: although this discussion does not argue directly for the Bayesian idea of rationality as expected utility maximization, it argues against a position that has been regarded as a contender in situations of strategic interaction.

1. INTRODUCTION

1.1.

Expected utility theory (the *Bayesian approach*) is a well-known theory of rational action.[1] Its basic idea appears in the *Port Royal Logic*: "In order to decide what we ought to do to obtain some good or avoid some harm, it is necessary to consider not only the good or harm in itself, but also the probability that it will or will not occur, and to view geometrically the proportions all these things have when taken together" (Arnauld and Nicole 1996, 273–4). The Bayesian-rational agent chooses an action that maximizes his expected utility. Is this approach always appropriate? Some say no. They invite us to consider *strategic* interaction, i.e., situations in which each of several agents is concerned to do as well as possible, and in which each agent's outcome depends on what everybody does. Given such interdependence, advocates of this position continue, the Bayesian approach gives the wrong answer. The rational recommendation, so they say, is not for each agent to assign probabilities to the possible actions of the others and to choose a utility-maximizing action. Rather, the agents should choose actions in *equilibrium*, i.e., actions such that no single agent gains from deviating if the others do not deviate.[2] Notable subscribers of this view are John Harsanyi and Robert Aumann.[3]

Synthese **124:** 361–384, 2000.

1.2.

I explore arguments for this view. The conception of equilibrium used is the *Nash Equilibrium*. Before I define this notion, I need to introduce the technical notion of a non-cooperative game: An n*-person non-cooperative game* is a set $\Gamma = (A_1, \ldots, A_n; u_1, \ldots, u_n)$, where the finite set A_i is agent i' strategy set and $u_i\colon A_1 \times \cdots \times A_n \to R$ is agent i's utility function. The sets A_i represent an agent's courses of action (i.e., strategies), which may be complicated sequences of moves. For instance, one strategy in chess would prescribe to an agent a reaction to every possible history of moves. The utility function for each agent is defined on the Cartesian product over all those strategy sets, which expresses the interdependence of outcomes. Such games are traditionally called *non-cooperative* because it is implicitly assumed that the agents cannot make binding agreements before choosing their strategies.[4] In two-person zero-sum games, $n = 2$ and $u_1 = u_2$. That is, these are games in which the gains of the one agent are always the losses of the other. A *Nash Equilibrium* is an n-tuple of strategies $a_1, \ldots, a_n$ such that $u_i(a_1, \ldots, a_i, \ldots, a_n) \geq u_i(a_1, \ldots, b_i, \ldots, a_n)$ for each i, where b_i is any strategy of agent i. That is, no agent can obtain higher utility as long as no other agent changes his or her strategy. Nash (1950) proves that if we use, instead of A_i, the set $P(A_i)$ of probability measures on A_i (i.e., allow for *randomized strategies*), then every game has at least one Nash Equilibrium.[5] These non-cooperative games model *one time-interactions*. In this essay, I shall largely ignore iterated games.

I argue that the arguments for Nash Equilibria as rational recommendations are either very limited in scope or fail. Presently, many economists do not endorse the equilibrium as rational solution concept to game situations.[6] *Rationalizability*, a consequence of common knowledge of rationality (cf. Bernheim (1984), Pearce (1984)), has gained much recognition.[7] But for one thing, the Nash Equilibrium has attracted enough attention as an allegedly rational recommendation for it to be worthwhile discussing these arguments. For another thing, the Bayesian position needs a view on this matter regardless of its acceptance among economists (even though, of course, rejecting the idea of equilibrium as a rational recommendation does not *by itself* support the Bayesian view). This inquiry goes against the advice of Aumann (1985), who argues that "a solution concept should be judged by its performance in the applications, by the quantity and quality of the relations that it engenders, not by 'armchair' philosophizing about its definition" (p. 42). Yet the concept of rationality is essential to the evaluation of human action. So it is crucial to appraise proposed criteria of rationality, and this question does not reduce to an assessment of success in applications.[8]

A word of caution about the terminology: Even though this essay is motivated by the idea that the Bayesian notion of rationality introduced above suffices for situations addressed by game theory, it is not this Bayesian notion of rationality that informs the discussion of the arguments for the Nash Equilibrium as a rational recommendation. Rather, this discussion is guided by an *intuitive* and *informal* notion of rationality appropriate in a context where the very suitability of a precise definition of rationality in action for a class of situations is at stake. Roughly speaking (and ignoring significant philosophical issues), according to this pre-theoretical notion of rationality, those actions are rational that best serve an agent's interest, whatever these interests are. Criticizing a proposed definition of rationality in action for a class of situations may then amount to discussing important cases in which this technical notion of rationality is too much at odds with the pre-theoretical idea of rationality.

2. THE THEORY OF GAMES AND ECONOMIC BEHAVIOR

2.1.

I first look at von Neumann and Morgenstern's (1944) argument for their *minimax solution* to zero-sum games. That argument does not support equilibrium recommendations in general, because minimax and equilibrium coincide only in two-person zero-sum games.[9] Still, the seminal role of this argument and the *Theory of Games and Economic Behavior* in general urges its discussion. Most of the *Theory of Games and Economic Behavior* is about cooperative games, i.e., deals with the formation of coalitions. The non-cooperative theory covers one-shot two-person zero-sum games only. Von Neumann and Morgenstern's rational recommendation is to choose a minimax strategy: If we represent the game in a matrix with the entries representing the row player's utilities, the latter is advised to choose a row with the highest minimum. (Recall that the gains of the first agent are the losses of the second agent, and vice versa.) The column player is advised to choose a column with the lowest maximum. So each is advised to maximize his security level. If mixed strategies are feasible, von Neumann and Morgenstern show that the outcome the row player obtains by pursuing his security strategy is the same as the outcome the column player loses by pursuing her strategy. This is the *Minimax-Theorem*:

$$\max_{s\in A_1}\min_{t\in A_2} u_1(s,t) = \min_{t\in A_2}\max_{s\in A_1} u_2(s,t).$$

Pairs of security-maximizing strategies (s, t) form equilibria, and all equilibria in zero-sum games are of this form. If we have two pairs (s, t) and (s', t') of such strategies, then (s, t') and (s', t) are equilibria and

$$\begin{aligned} u_1(s,t) &= u_1(s,t') = u_1(s',t) = u_1(s',t') = -u_2(s,t) \\ &= -u_2(s,t') = -u_2(s',t) = -u_2(s',t'). \end{aligned}$$

This number $u_1(s, t)$ is called the *value* of the game.

2.2.

Let us reconstruct von Neumann and Morgenstern's argument for security-maximizing strategies as *rational recommendations*. Suppose Γ is a two-person zero-sum game. The obstacle to providing rational advice is that neither player knows what the other is doing even though his strategy choice depends on what he thinks the other does. Von Neumann and Morgenstern introduce two auxiliary games to solve this problem. The minorant game Γ_1 equals Γ except that player 1 moves first, player 2 knows 1's action when she, 2, moves, and that all this is common knowledge. The *majorant game* Γ_2 equals Γ but player 2 moves first, player 1 knows 2's action when he, 1, chooses and all this is common knowledge.[10] In Γ_1 and Γ_2, rational advice is available. In Γ_1, player 1, moving first, knows that when player 2 moves, she will hurt him as much as she can (zero-sum game). So the rational advice to player 1 is to maximize his minimum. Mutatis mutandis for Γ_2. Define a value v_1 for the minorant game and a value v_2 for the majorant game:

$$v_1 = \max_{s \in S_1} \min_{t \in S_2} u_1(s,t) \text{ and}$$

$$v_2 = \min_{t \in S_2} \max_{s \in S_1} u_2(s,t).$$

As is easy to see, $v_1 \leq v_2$. Von Neumann and Morgenstern claim that if there is a value v for Γ, it must be $v_1 \leq v \leq v_2$ (p. 106). For Γ_1 describes player 1's most favorable situation and Γ_2 describes player 2's most favorable situation. So these two outcomes should constitute limits for what to expect for Γ. Von Neumann and Morgenstern want to continue without making any assumptions about "the players' intellects" (p. 111). They point out that "a satisfactory theory can exist only if we are able to harmonize the two extremes Γ_1 and Γ_2, strategies of player 1 'found out' and strategies of player 2 'found out' ".[11] Next, then, they introduce mixed strategies and define a minorant game and a majorant game for the mixed strategy case. For these games, we find values v_1' and v_2' analogous to v_1 and v_2. They demonstrate that $v_1 \leq v_1' \leq v_2' \leq v_2$. The Minimax theorem shows that $v_1' \leq v_2'$ does not happen. The "good way" (p. 159) to play for the players is to choose a security strategy.

This argument is intriguing but unsatisfactory. Security-maximizing strategies in Γ_1 and Γ_2 are rational because one player reacts at stage 2 to what the other one did at stage 1. Yet what does this imply for Γ? One may argue as follows: If Γ were played with the goal to make things maximally favorable for 1, v_1 would be forthcoming. If Γ were played with the goal to make things maximally favorable for 2, v_2 would be forthcoming. So if Γ is *not* played with either goal, there should be a value *between* them. The admission of mixed strategies then leads to v. This argument assumes that *there is* a value for Γ, i.e., a unique rational recommendation. Yet although Γ_1 and Γ_2 make for unique rational advice, this does not imply that there is such advice for Γ. For in Γ_1 and Γ_2, the agents move one after the other, whereas in Γ, they move simultaneously. So in Γ we *abandon* the very assumption that makes unique rational advice available.[12] Unique rational recommendations can be given in situations suitable for this kind of advice. Γ_1 and Γ_2 provide examples, but Γ does not.[13]

Discussing zero-sum games, Luce and Raiffa (1957) say: "What should he (player 1) do? Game Theory does not prescribe what he should do! It does point out that player 1 can guarantee himself 4 by selecting ... and that no other choice has as high a guarantee, but what 1 should do the theory is careful to avoid saying" (pp. 60–3). A little later: "[I]t is crucial that the social scientist recognize that game theory is not descriptive, but rather (conditionally) normative. It states neither how people do behave nor how they should behave in an absolute sense, but how they should behave if they wish to achieve certain ends" (p. 63). This is more modest than the *Theory of Games and Economic Behavior*. Yet what are the ends conditional on which minimax play is the most appropriate course of action? The most plausible candidate is that the end is to maximize one's security. But then Luce and Raiffa (1957) do not tell us anything new: if you want to maximize your security level, use minimax strategies. But that is what minimax strategies are designed for.[14] Nothing improves by pursuing the more modest line. I conclude that von Neumann and Morgenstern fail to show that minimax strategies are the unique rational recommendation for two-person zero-sum games. Yet this does not tell us anything about arguments for equilibria in general, since equilibria and minimax strategies part company in games other than two-person zero-sum games. I proceed to arguments for equilibria proper.

3. THE ARGUMENT FROM SELF-ENFORCING AGREEMENTS

3.1.

This influential argument runs as follows: Nash Equilibria are recommended by being the only strategy combinations on which the players could make self-enforcing agreements, i.e., agreements that each has reason to respect, even without external enforcement mechanisms.[15] Can all Nash Equilibria be generated by self-enforcing agreements? The answer is negative, as is well-known (cf. Kreps (1990, 32)). Consider the following game:

4, 4	4, 4	0, 0
4, 4	4, 4	0, 0
0, 0	0, 0	1, 1

Playing the third row and the third column is an equilibrium. But agents may have reasons to deviate from their respective strategies and hope that the other one does likewise. They would reach a Pareto-superior Nash Equilibrium. Yet nothing depends on the fact that what the players may deviate to is an equilibrium. Consider the next game:

4, 6	5, 4	0, 0
5, 7	4, 8	0, 0
0, 0	0, 0	1, 1

The lower right corner is the only Nash Equilibrium in pure strategies. Still, there is reason for the agents to deviate and to expect the other to do likewise. Yet these two examples are not very interesting. For neither equilibrium would be agreed upon by agents who are in their right minds. Next (cf. Aumann (1990)) we encounter an equilibrium agents may abandon for a Pareto-*inferior* one:

9, 9	0, 8
8, 0	7, 7

This game has two equilibria in pure strategies: (Top, Left) and (Bottom, Right). (Top, Left) is Pareto-superior to (Bottom, Right). (Bottom, Right), however, is risk-dominant (in a sense that should become clear enough in the discussion). Suppose the agents meet before the game. Since (Top, Left) is Pareto-superior, they have good reason to agree on playing these strategies. Suppose before the game they both reflect on the agreement. Focus on the row player. He may think that the column player might suspect him of not trusting her. He may think that she may play Right to ensure at least 7 for herself. To protect himself he defects to Bottom. Reaffirming this decision, the row player may think that the column player shares these thoughts. That is, the row player thinks that the column player thinks that he does not trust her and that he thinks that she may defect from the agreement. Therefore, the row player has good reason to suspect that the column player defects from the agreement, which gives him all the more reason to defect himself. As Aumann (1990) points out, the agreement fails to convey information to either player. In non-cooperative games, the players cannot make commitments that are not in their interest to carry out later. Such agreements merely express what they want the *other* players to do. But in this game, the row player knows already that the column player wants him to play Top, since her payoff is higher then no matter what she does herself. Similarly, the column player knows that the row player wants her to play Left.

The game models Rousseau's stag hunt from the *Discours sur l'Inegalité*. Two hunters cooperating will catch the stag, but they must follow different paths and lose sight of each other. Each sees a hare crossing his path. Catching the hare ensures a meal for one person, but dooms the stag hunt. Each may doubt the other's trustworthiness and catch the hare. Or each might think that the other thinks him untrustworthy. Or each may think that the other thinks that he does not trust him, which is why the other person has a reason to deviate, which in turn provides an incentive for him to pursue the hare. "Higher levels of doubt" may have the same effect as plain old distrust – and are easier to maintain. Upon reflection I see that it is paranoid not to trust you. I *do* trust you – but I worry that you may think of me in ways that make you lose trust in me. Or maybe you suspect that something like that is on my mind? Such considerations are not far-fetched. They frequently prompt anxieties in everyday life. Hobbes, in the *Leviathan*, discusses such phenomena under the heading of *anticipation*. Describing the state of nature and explaining why in this state life is "solitary, poor, nasty, brutish, and short", Hobbes claims, early in chapter XIII, that "there is no way for any man to secure himself, so reasonable, as Anticipation".[16]

3.2.

I have argued that not all Nash Equilibria can be obtained through self-enforcing agreements. It may still be true that *only* Nash Equilibria can be so obtained, which is a view expressed by Kreps (1990). Are there self-enforcing agreements on strategies other than Nash Equilibria? Consider this game:

0, 0	4, 2
2, 4	3, 3

There are two equilibria in pure strategies ((Top, Right), (Bottom, Left)). Suppose the agents discuss the game beforehand. (Bottom, Right) suggests itself as a compromise, and they settle for it. When it is time to honor the commitment, the row player may think that deviating profits him, and the column player may think so, too. However, they risk ending up with nothing if they deviate. Sticking to the agreed-upon strategy means sticking to their security strategy, and the fact that they agreed on these strategies reassures them that there is a point to doing so. One may object that an agreement is not *self-enforcing* only because there is some point to honoring it. It is self-enforcing only if there is no serious consideration about deviating. In this case, there is a serious reason for deviating. For each agent is better off by choosing his other strategy as long as the other one does not do so. So what *are* self-enforcing agreements?

The straightforward answer is that self-enforcing agreements are those in which nobody gains by deviating. This is a bad answer. For we cannot use the concept of self-enforcing agreement in arguments for Nash Equilibria as rational recommendations if the two concepts are so closely related in their definitions. Instead, one could say this: Self-enforcing agreements are those from which nobody has any reason to deviate if nobody else does. But this is still too close to the definition of Nash Equilibria. The notion of *reason* does not bring anything new into the definition: this just amounts to saying that there is no gain from deviating by oneself. We have to remove the conditioning subclause to obtain a concept of self-enforcing agreement more detached from Nash Equilibria. A self-enforcing agreement is one that provides incentives for the agents to stick to it even in the absence of external enforcement mechanisms. Yet this definition renders agreements on non-equilibrium strategies self-enforcing, e.g., an agreement on (3, 3) in the preceding game. So a useful definition of self-enforcing agreements covers more than just equilibria.[17]

3.3.

I have established, then, that being the subject of self-enforcing agreements is neither necessary nor sufficient for being a Nash Equilibrium.

Another question arises: Why does it matter for assessing the rationality of a solution that it could be the outcome of a self-enforcing agreement? An argument to the contrary is that players do not always have an interest in coordinating their actions. Consider the following game, taken from Farrell (1988):

−2, −2	1, −3	1, −3
−3, 1	2, −2	−2, 2
3, 1	−2, 2	2, −2

The unique Nash equilibrium is the upper left corner, and every outcome is rationalizable. As Farrell points out, whatever one agent thinks the other will do if this equilibrium is not suggested in communication, she can only lose by suggesting it. So it is not suggested. Communication is of no interest for the agents. So insisting on a criterion of rational agency that dwells on what agreements would be self-enforcing does not cover this and similar situations. As another example, take the famous Holmes–Moriarty scenario:[18] What relevance could it have that a Nash Equilibrium is the only outcome of a self-enforcing agreement? More generally, the range of non-cooperative games is so large that it is dubious that in all or even most of them self-enforcing agreements are appealing enough on rational grounds to recommend whatever has been so agreed upon and nothing else as a rational recommendation. Analogous points apply to justifications of Nash Equilibria as solutions that a referee would have reason to suggest; or against the argument that only Nash Equilibrium strategies will be implemented after they have been publicly announced.[19] All of these arguments employ a test for rationality that goes unsupported. To be sure, arguments of this kind *illuminate properties* of Nash Equilibria. We learn under what circumstances they would occur or would be rational recommendations. Yet these arguments fail to show that Nash Equilibria have a special role to play as rational recommendations *sui generis*.

3.4.

One last point: It is hard to see how the argument from self-enforcing agreements accounts for mixed equilibria. Recall that in mixed equilibria all strategies with positive probability are best replies to the other agent's strategy. So once a player's random mechanism has assigned an action to her, she might as well do something else. Even though the mixed strategies might have constituted a self-enforcing agreement *before* the mechanism made its assignment, it is hard to see what argument a player should have to stick to the agreement after the assignment is made.

The upshot of this discussion is that the argument from self-enforcing agreements fails. Not all Nash Equilibria can be the outcome of self-enforcing agreements, and not all self-enforcing agreements lead to Nash Equilibria, once we have developed a concept of self-enforcing agreements that is sufficiently detached from the concept of Nash Equilibrium to be helpful. Finally, it is not required of rational recommendations for game situations that they pass this “test” in the first place.

4. THE ABSENCE-OF-PROBABILITIES ARGUMENT

4.1.

This argument claims that in strategic interaction, probabilities are unavailable. For each agent’s probabilities would depend on what he thinks the others will do, but he also would know that what the others will do depends on what they think his probabilities are. There will not be sufficient reason to settle for any probabilities. In such situations, rational advice is to realize an equilibrium, which does not depend on probabilities.[20] Three questions arise: First, are there really situations in which no probabilities can be assigned? Second, if there are any such situations, are strategic interactions among them? Third, if strategic interactions are of that type, does this speak for Nash Equilibria as rational recommendations? It suffices to discuss the second (sketchily) and third question.

As for the second question, traditional strategic thinking did not find difficulties in using probabilities. Clausewitz (1972, 199) points out that war parties assess *probabilistically* what the other parties will do, using what they know about their character, institutions, and situation. Clausewitz seemingly felt comfortable practicing the advice of the *Port Royal Logic*. His reflections on strategy exclude higher levels of reasoning, which might possibly invalidate assignments of probabilities. Likewise, an orator in Thucydides’ *Peloponnesian War* advises his compatriots to design their own strategies in response to “that which wise men and men of great experience . . . are likely to do”.[21] Again, no mention of higher levels of second-guessing. Seemingly, either such higher levels were perceived as obscure or as unnecessary for decision making.

My main response to the absence-of-probabilities argument is a reply to the third question. There are several well-known rules for decision making when probabilities are not available, e.g., the minimax regret rule, the maximin rule, the optimism-pessimism rule. These rules do not always offer the same advice (cf. Resnik (1987, 38 for examples), and we lack conclusive arguments for any of them. Similarly, it is hard to see why choosing a Nash Equilibrium strategy is, in general, superior to choosing a strategy recommended by another rule. Nothing in the absence of probabilities by itself calls for Nash Equilibrium strategies. What is missing is a reason why the agents should conceive of non-cooperative games as situations that require a *joint* solution of the kind envisaged by Nash Equilibria (cf. also Section 6). So the argument from the absence of probabilities fails: Even if we grant that in situations of strategic interaction no probabilities can reasonably be assigned, this does not argue for Nash Equilibria as rational recommendations.

5. THE TRANSPARENCY-OF-REASON-ARGUMENT

5.1.

This argument runs as follows: Rational agents figure out each other's thought processes. Therefore, whenever one of them assigns probabilities, the others predict them and assess their probabilities accordingly. The original agent, in turn, figures this out and adjusts his probabilities. This ends only when the agents reach a Nash Equilibrium.[22]

What does it mean that agents figure out each other's strategies? Suppose the agents can *read each other's thoughts*. Would they arrive at an equilibrium? They would not. Suppose we have only two agents. Suppose they reach an equilibrium by reading each other's thoughts and responding in a maximizing way. The following may happen: One agent judges deviating reasonable if the other one deviates as well. The other one figures this out. Suppose mutual deviation is profitable for him as well. Since he knows that the first agent would deviate if he did, he deviates. In this way, two agents transparent to each other could abandon an equilibrium. The point is that a Nash Equilibrium, by definition, only discourages *uni*-lateral deviation. This objection could be met by pointing out that the process of mutual out-guessing would not stop at a joint deviation. If either one of those agents deviating could improve his outcome by using a strategy different from the one he envisaged himself deviating with once he has figured out that other person would deviate with him, he would choose that other strategy. The other agent, however, would figure *that* out, and they would move towards an equilibrium. They would stop at an equilibrium that is Pareto-superior to the others, once they have gone through all possible deviation scenarios. So at least this argument works for those Pareto-superior equilibria.

However, this rescue strategy fails. So far this discussion has assumed that the agents deliberate in a simple way, one thought at a time, and do not look far before they leap. What if the agents have more elaborate ways of thinking about each other, i.e., thoughts such as "Suppose I tentatively considered to deviate; then my opponent would figure this out and respond by doing X. But then I would respond by pretending to consider Y, whereas really I would consider Z, etc.". Would the other agent figure this out step by step, or would she figure out the whole train of thought at once, and just when would her own *having-figured-it-out* enter into the first agent's thoughts? As soon as we deviate from the simplistic picture that the agents have one (simple) thought at a time without looking ahead we cannot even make sense of this "figuring-out" any more. We cannot say what the units of thought (so to speak) are that are figured out, and just when to think of one such unit to be figured out by the other agent, so that then the first agent can integrate his having been figured out into his next unit of thought. Clearly, the simplistic way of figuring each other out "thought-by-thought" is not what we should assume. For it means assuming immense mental capacities and then thinking of them in a naive way. That pushes us towards this more elaborate way of interpreting their figuring each other out, and that way is hard to make sense of.

Yet even if we could establish that transparent people reach a Nash Equilibrium, we would not have gained much. For such transparency is not a feature of *rationality*. At most we could show something for people with this remarkable property, without having made a contribution to our understanding of rationality. This naive version of the transparency-of-reason argument is seriously defective in various ways.[23]

5.2.

Bacharach (1987) presents another version of the transparency-of-reason argument. Restricting himself to two-person games, Bacharach uses a language of first-order predicate logic with epistemic operators. Classes of games are described axiomatically in this language, i.e., in terms of theories in the sense of first-order predicate logic. In such a framework, solution concepts appear as pairs of one-place predicates (one for each agent). A pair of predicates is properly regarded as a solution concept in those theories that include as theorems six sentences (three for each agent) containing those predicates. The first of these sentences for each agent says that an action is satisfactory for an agent if and only if this person's predicate is true of it. The second says that the agent recognizes which actions satisfy his solution predicate and which ones satisfy the other agent's solution predicate. And the third says that there is only one such action for each agent.

This characterization of solution concepts in this language of epistemic logic then interacts with a rule of inference according to which theorems of the theory are always known to players. Using this machinery, Bacharach proves a theorem he calls the *Transparency-of-Reason Theorem:* If some action *a* is the action such that agent 1(2) knows that it is the uniquely satisfactory thing to do for him, then agent 2(1) knows all this. In a next step Bacharach shows that only Nash Equilibria (properly represented in his system) can be solutions in the sense defined above, using the definition of a solution concept, the Transparency-of-Reason Theorem, and a principle called the best-response principle. The best-response principle says this: If agent 1(2) knows what agent 2(1) does, then he does not consider any action satisfiable that is not a best response.

Since Bacharach's argument is not presented in a dynamic setting, it is hard to find conceptual space for the objection that it makes no sense to think of the agents as figuring out each other. However, we are now back with the worries raised against the von Neumann–Morgenstern argument: Why is it appropriate to assume that there is a unique rational recommendation in game situations? The burden of proof is on those who want to classify actions as irrational (or at least as not rational). For the concept of rationality loses its significance if not applied cautiously; it should be used restrictedly rather than restrictively. Again, it seems appropriate to say that we learn something about the properties of equilibria, but not that they turn out to be rational recommendations in cooperative games.

5.3.

Harsanyi (1982) gives what can be read as a more modest version of the transparency-of-reason argument. Harsanyi (1982) replies to Kadane and Larkey (1982) who claim that "in a single play, all aspects of [the player's] opinion except his opinion about his opponent's behavior are irrelevant" (p. 116). Higher levels of second-guessing do not play any role. Harsanyi responds by pointing out that this amounts to throwing away valuable information, i.e., the information that there is common knowledge that all agents are rational. Moreover, he asserts that the most interesting problem of game theory is "how to translate the intuitive assumption of mutually expected rationality into mathematically precise behavioral terms (solution concepts)" (p. 121). I regard this as a version of the transparency of reason argument because it appeals to common knowledge of rationality. It is a modest version because it does not involve agents who can figure out each other's strategies. However, it is well known today that common knowledge of rationality does not lead to any equilibrium concept. It merely implies iterated elimination of no-best-response strategies (cf. Rubinstein and Osborne (1994), chapter 4). The strategies not eliminated in this process are those for which there is a story of the kind "I play this strategy because it is a best reply to what I think my opponents will do, and I think they will play those strategies because they are best replies to what I take to be their conjectures about the other players, etc.". The appeal to common knowledge of rationality cannot be used to support equilibrium reasoning.

Harsanyi would reject this reasoning. Harsanyi (1977, 10) points out that "the basic weakness of traditional game theory has been that in defining rational behavior in game situations it has tried to restrict itself to using rationality postulates which in their logical content do not go significantly beyond the rationality postulates of individual decision theory". Later then, pp. 116–8, he introduces additional postulates that lead up to equilibrium recommendations. One of these postulates requires that each player use a best reply to what the others do, and from here the equilibrium is within reach.[24] However, this move is not helpful at this stage of my argument, because Harsanyi assumes that the agents face a *joint problem* and conceive of themselves in such a way. But that is not always so. The agents deal with a problem of interdependence, certainly; but that does not imply that they face a *joint* problem. Thus Harsanyi's argument against Kadane and Larkey fails, and so does the modest version of the argument from the transparency of reason.

6. THE ARGUMENT FROM REGRET

6.1.

This argument is the following: If the agents choose any strategy combination other than a Nash Equilibrium, at least one of them regrets her choice once she sees what everyone else does. Another actions gives her higher utility. She has lost, as Lewis (1969, 8) puts it, "through lack of foreknowledge". A reply to this argument is that the occurrence of regret after all relevant facts are known does not entail that the original choice was irrational. Yet this reply can be met by pointing out that for an agent to choose a non-equilibrium strategy is choosing a strategy for which it is *certain* that some person will have regrets. However, the fact that *some* agent will have regrets is certain to motivate the agents only if they conceive of game situations as *joint problems.* To save the argument from regret, we need to provide a reason why they should so conceive of games.[25]

This task is hopeless. The agents do not in general aspire at a joint solution because, being involved in a non-cooperative game, they do not always care about their opponents, or think of themselves *as a team*. Thus it is misguided that Sugden (1993) should scold game theory for not providing means to argue that the agents should avoid a Pareto-inferior equilibrium (as they would if they thought of themselves as a team). For teams would not aim at equilibria in the first place. As a team, agents try jointly to obtain as much as possible and distribute the gain among themselves. Non-cooperative game theory does not (and never intended to) provide guidance to teams for their internal distribution problems. Thus if there is any argument for the agents to look at the situation as a joint problem it can only draw on each person's own interest. Yet the appeal to rationality alone does not prompt them so to conceive of the situation. For even common knowledge of rationality only entails that no strategy is used that fails to survive iterated elimination of strategies that are not best responses to some strategy combination of the opponents, i.e., the rationalizable strategies. To execute the iterated elimination, the agents do not have to think of the problem as a joint one. The argument from regret fails. By now it should also be clear that one pervasive problem with arguments for Nash Equilibria as rational recommendations is the assumption that the agents conceive of their situation as a joint problem. This assumption is misguided because the setting is by definition non-cooperative.

7. THE ARGUMENT FROM CORRELATED EQUILIBRIUM

7.1.

The next argument appears in Aumann (1987b). If Bayesian rational agents use a common prior probability measure, their strategies form a correlated equilibrium. So Bayesian rationality *all by itself* leads to a certain type of equilibrium – not the Nash Equilibrium, but the more general correlated equilibrium. This requires elaboration. Readers familiar with Aumann (1987b) may proceed to 7.2. I begin with terminology. Aumann's concept of Bayesian rationality differs from the one presented in the introduction; I define the current notion as we go along. The idea of a correlated equilibrium is simple.[26] Consider the game of chicken:

6, 6	2, 7
7, 2	0, 0

This game has three Nash equilibria with payoffs (2, 7), (7, 2), and (4 2/3, 4 2/3). The players may obtain higher payoffs if they could randomize over the joint actions space. E.g., the following distribution of probabilities over strategy pairs is impossible if the players randomize independently, but feasible if they use a joint randomization mechanism:

1/3	1/3
1/3	0

The payoff is (5, 5). The point of correlated equilibria is that the players act in accordance with a randomization mechanism defined on their joint strategy space. Imagine a referee observing outcomes of a random process and announcing to each player what he has to do according to those outcomes. The agents only know the probabilities over the joint strategies. Formally, we have the following model: Let G be an n-person game, and denote by S_i, $1 \leq i \leq n$ agent i's the strategy space. S is the Cartesian product over the sets S_i. Let u_i: $S \to R$ be the utility function for agent i. A *correlated strategy n-tuple* is a random variable from a finite probability space Γ into S. Such a correlated strategy n-tuple captures joint randomization. Define as the *distribution* of a correlated strategy n-tuple f the function that assigns to each n-tuple s of actions the number Prob $f^{-1}(s)$. The function f can be written as $f = (f_1, \ldots, f_n)$, and we set $(f_1, g_i) = (f_1, \ldots, g_i, \ldots, f_n)$. E denotes expectation. We can then define a correlated equilibrium: A correlated strategy n-tuple in G is a *correlated equilibrium* if and only if $Eu_i(f) \geq Eu_i(f_i, g_i)$ for all i and functions g_i which are functions of f_i. Every convex combination of Nash Equilibria is a correlated equilibrium, but there might be correlated equilibria which are outside the convex hull of Nash Equilibria (the one given in the *chicken* example above is outside the convex hull of Nash Equilibria). For the sake of concreteness, you may think of the initial idea introduced in connection with the game of chicken.

To establish Aumann's (1987b) main result, we use the following model: Again there is an n-person game G. In addition, there is a finite space of states of the world Ω, for each i a probability measure p_i on Ω, and for each player a partition P_i, representing i's information partition. One state of the world ω is the true state, and for every i there is a $P_i \in P_i$ such that $\omega \in P_i$. This P_i represents i's knowledge. Moreover, there is a common prior assumption, i.e., there is a probability measure p on Ω such that $p_i = p$ for all i. The agents' probabilities are posterior probabilities, i.e., p_i conditional on the agent's respective information. So the common prior assumption does not imply that all the agents share the same probabilities, but that differences in probabilities are due to differences in information *only*. Finally, define an agent as *Bayes rational at* ω if $E(u_i(s)|P_i(\omega)) \geq E(u_i(s_{-i}, s_i)|P_i(\omega))$. To explain this expression: $u_i(s)$ is a random variable. For any random variable f, $E(f|P_i)$ is a function that assigns to each ω the expected value of f conditional on the information cell P_i that ω is contained in.[27] We can then state Aumann's (1987b) main theorem:

> If each player is Bayes rational at each state of the world and if there is a common prior, then the distribution of the action n-tuple s is a correlated equilibrium distribution.

7.2.

Let us consider a number of worries about the philosophical import of this theorem. Levi (1997) raises an objection against the assumption that the agents know that they are rational, which is a fortiori an objection against the assumption of common knowledge of rationality.[28] He points out that an agent who wants to use principles of rational choice self-critically, i.e., who wants to deliberate, cannot predict, and therefore cannot know, that he will act rationally. Otherwise deliberation would be vacuous, since the outcome is determined when the relevant parameters of the choice situation are available. But, so the objection goes on, game theory in general and Aumann (1987b) in particular conceive of agents as deliberating about what they are going to do in view of their own and the others' utilities and their relevant beliefs. But then game theory cannot also assume that agents *know* that they are rational. Thinking of agents both as deliberators and as knowing their rationality leads to an incoherence.

Dekel and Gul (1997) seem to indicate a way around this objection. They show that Aumann's theorem does not depend on the assumption of common knowledge or even knowledge of rationality. So Levi's objection does not seem to arise. They point out that the theorem requires the support of the prior p to be a subset of the set [*rationality*] $\cap$ [u], where [*rationality*] is the set of all states of the world where each agent is an expected-utility maximizer and [u] is the set where the agents have utility functions u_i respectively. I.e., what is required is that rationality and the game played are certain. No knowledge assumption occurs. Under certain circumstances (e.g., finiteness of the model) this event coincides with the event of common knowledge of rationality. Yet that is not in general true. In general, the common knowledge assumption is stronger. Dekel and Gul prove the following theorem, generalizing Aumann's theorem without the common knowledge assumption (p. 134):

> Consider a sequence of models which differ only in the common prior: $\{\Omega, F_i, p_n, s, u\}_{n=1}^{\infty}$. Assume that u is bounded, i.e., $u(\omega)(s) < b \in R$ for any ω and all s. Let E^n be the event that the payoffs are u and the players are rational in model n. If $p^n(E^n) \to 1$, then the limit of any convergent subsequence of the distribution on actions induced by p^n and s is a correlated equilibrium distribution.

Again, no knowledge assumption occurs. Therefore, Levi's criticism does not seem to arise. However, Levi can restate his objection even in the model offered by Dekel and Gul. For he appeals to Aumann's assumption of common knowledge of rationality only in order to take note of the fact that it implies that each decision maker is *certain* that he will choose rationally. Yet that much is also entailed by the assumptions used by Dekel and Gul, since the support of the prior probability is a subset of set [*rationality*] $\cap$ [u]. So a successful rebuttal of Levi's objection would require a direct engagement with his worries about the idea that the agents are certain of their own rationality. However, for the purposes of this essay, pursuing this point any further would be disproportionate, since nothing else turns on it.

Another objection raised against the philosophical relevance of this theorem concerns its dependence on the common prior assumption.[29] Aumann (1987b) and other economists refer to Savage (1954) for a theory of subjective probabilities. Morris (1995, 233) sees a tension in using the common prior assumption in Savage models and therefore in all models that draw on Savage's account of probabilities. His argument is that such a prior would be exogenous to the model under scrutiny (i.e., not itself derived from data in that model) and could not itself be interpreted as a personalistic probability measure. So assuming a common prior is acknowledging a non-personalistic interpretation of probability, which Savage rejects. It is unclear whether Morris claims that the addition of a common prior to Savage's model leads to inconsistency. This would be wrong. For there is nothing in the fact that probabilities are elicited from preference rankings (which is what makes them personalistic in the Savage model) that prevents them from also being derived from a common prior. Similarly, the fact that a common prior would be exogenous to the problem does not imply that the prior is non-personalistic. This common prior could be a common *ur*-prior in the Carnapian sense, shared by all rational agents. So this objection against the common prior assumption relies on too strong an idea of personalistic probabilities.

7.3.

There is, however, a more serious objection against the significance of Aumann's result. The model does not simply assume that each person chooses an action maximizing his expected utility. Rather, everybody maximizes his expected utility *given what the others do.* An action that is still rational once an agent has made up his mind is called *ratifiable* in the literature.[30] The situation described in the given model is one of *joint ratifiability*: Each agent's action still maximizes her expected utility after *all* of them have made up their minds. Therefore, the assumption itself makes sure that we find an equilibrium. Now we also see more clearly the difference between the concept of Bayesian rationality that I introduced in 1.1, and the concept of Bayesian rationality used in Aumann (1987b). In 1.1, we encountered an *ex ante* notion of rationality. I.e., each agent's rationality consists in choosing an action independently of the others that maximizes his expected utility. Whether or not this is a situation in which each of these actions still maximizes the respective agent's expected utility *given* what all of them have chosen remains to be seen. As opposed to this, Aumann's (1987b) concept of rationality is an *ex post* notion of Bayesian rationality. His main result does not hold for the *ex ante* notion, but only for the *ex post* notion. The scope of this argument is therefore quite narrow.

Levi criticizes Aumann's argument for taking ratifiability as a requirement of rationality in the first place. Whether ratifiability is a criterion of rationality is not my concern. Surely, as Jeffrey (1983, 18), put it, the adoption of this criterion "modifies the Bayesian maxim in a way that makes no difference in the usual, straightforward sorts of problems". So whether or not we adopt the principle of ratifiability as a principle of rationality makes a difference only for unusual cases, and we would not want too much weight attached to rejecting such an addition to a theory of rationality. Rather, my point is that *joint* ratifiability is extremely demanding.[31] In particular, and quite naturally, all agents may act rationally and not end up in a situation of joint ratifiability.

8. CONCLUSION

We have reviewed a number of influential arguments for the rationality of Nash Equilibria. None of them could establish for more than a limited class of strategic interactions that Nash Equilibria are the rational recommendation in one-time interaction. Most of them failed completely. Two pervasive objections (in addition to problems peculiar to specific arguments) are that, first, it is assumed without justification that there simply must be a distinguished kind of recommendation for strategic interaction; and that second, it is assumed without justification that agents conceive of game situations as joint problems. That does not exclude that Nash Equilibria have some role to play in repeated interactions of rational agents, for instance as the result of rational learning (cf. for instance Kalai and Lehrer (1993)), or that Nash Equilibrium or any other type of equilibrium should occupy a prominent role elsewhere. However, whatever else it is, Nash Equilibrium is no rational recommendation for one-time interactions.[32]

NOTES

[1] Cf. Fishburn (1981) for an overview of various theories of expected utility maximization. "Bayesians" in some sense are, for instance, Savage (1954) and Jeffrey (1983, 1992).

[2] If the agents choose strategies forming an equilibrium and assign probability 1 to the other agents' choosing their equilibrium strategies, they all maximize expected utility. But the converse is false.

[3] Aumann (1985, 43), points out that the concept of (Nash) equilibrium is the embodiment of the idea that economic agents are rational and act simultaneously to maximize their utility. In addition, the concept of Nash equilibrium expresses that economic agents act in accordance with their incentives. Harsanyi is a Bayesian in advocating subjective probabilities. However, he is not a Bayesian in the sense introduced here, where each agent assigns probabilities to the possible actions of the others and chooses a utility-maximizing act. Harsanyi (1977, 11), presents a general theory of rational behavior, which consists of individual decision theory and a theory of rational behavior in a social setting. This latter theory consists of game theory and ethics; the rational recommendation for game theory is to go for an equilibrium, with the exception of situations in which the agent cannot rationally expect to gain more than what his security strategy ensures him (p. 116).

[4] This point is not part of the definition, but cf. see Section 3 on how the impossibility of certain kinds of cooperation emerges from this definition.

[5] For a more detailed exposition, cf. for instance Rubinstein and Osborne (1994). I do not discuss equilibrium refinements, because introducing them would not change much.

[6] For instance, Kreps (1990, 31) writes: "Unless a given game has a self-evident way to play, self-evident to the players, the notion of a Nash-Equilibrium has no particular claim upon our attention".

[7] Roughly speaking, rationalizable strategies are those that survive iterated domination of strictly dominated strategies. That is, we begin with a game Γ and remove in each strategy set those strategies that the player has never reason to choose because there is a strategy that always gives him higher payoff no matter what the others do. Doing this in each strategy set gives us a new game, and again we remove all these dominated strategies, etc. Cf. Rubinstein and Osborne (1994, Chap. 4), for more extensive treatment.

[8] Aumann (1987a), however, calls a game theory a "unified theory for the rational side of social science" and should thus grant that solution concepts of game theory need to be assessed as criteria of rationality.

[9] These two concepts are very different. As Aumann and Maschler (1972) point out, in minimax reasoning the agent plays for himself rather than against the others, whereas in

equilibrium reasoning the agents plays against the others rather than for himself. To see this, consider that an agent's minimax strategies depend only on his own utilities, whereas his equilibrium strategies depend on the other person's utilities (i.e., we can change the agent's utility function significantly without changing his equilibrium strategy, but not so for his maximin strategy).

[10] Cf. chapter 14 of their book.

[11] Von Neumann and Morgenstern assume that there is no difference between a player's strategy having been found out and this player being the first to move, where this move is known to the second player when she moves.

[12] Von Neumann and Morgenstern point out on p. 104: "[W]e have the hope that the definition of the value of a play may be used in the same form for other games as well – in particular for the game Γ – which, as we know, occupies a middle position between Γ_1 and Γ_2. This hope applies, of course, only to the concept of value itself, but not to the reasoning which leads to it; those were specific to Γ_1 and Γ_2, indeed different for Γ_1 and for Γ_2, and altogether impracticable for Γ itself". Yet conceiving of Γ as occupying such a middle ground is misleading.

[13] Ellsberg (1954) scolds this argument for equating rationality with defensiveness.

[14] Nozick (1990) comments on this in the same way; he is aware of Ellsberg (1954) and takes his argument to be decisive criticism against the von Neumann–Morgenstern argument. Resnik (1987, 130), however, accepts the von Neumann–Morgenstern reasoning without any discussion.

[15] The arguments discussed in this essay are part of the game theory folklore and normally do not have a locus classicus. I am therefore not concerned with documenting sources where they have been used.

[16] "Anticipation" is also a motive in Thucydides' *Peloponnesian War*, which Hobbes (!) translated into English; cf. Book 6, sections 18 and 38.

[17] The outcome (3, 3) constitutes a *focal point* of this game in Schelling's (1960) sense; this case therefore rules out a possible focal point argument that one might want to make in favor of Nash Equilibria. Sometimes Nash Equilibria are focal points, sometimes they are not. And sometimes focal points are Nash Equilibria, and sometimes they are not; so there is no argument from focal points for Nash Equilibria.

[18] This story is analyzed, for instance, in von Neumann and Morgenstern (1944, 176–7). Holmes and Watson embark on a train from London to Dover in order to escape to the Continent. As the train leaves, Professor Moriarty, determined to kill Holmes and Watson, appears on the platform. It is common knowledge among all of them that Moriarty will hire a special train in order to catch up with them. The regular train only stops once on its way from London to Dover, in Canterbury. Where should Holmes and Watson get off, and where should Moriarty get off? The way von Neumann and Morgenstern have set up the game, Holmes–Watson should get off at Canterbury with a probability of 0.6, whereas Moriarty should get off there with a probability of 0.4.

[19] An early version of such an argument from publicity was championed by Luce and Raiffa (1957, 63, and similarly on p. 173): "It seems plausible that, if a theory offers α_{i0} and β_{i0} as suitable strategies, the mere knowledge of the theory should not cause either of the players to change his choice: just because the theory suggests β_{j0} to player 2 should not be grounds for player 1 to choose a strategy different from α_{i0}; similarly, the theoretical prescription of α_{i0} should not lead player 2 to select a strategy different from β_{j0}". This suggestion seems to rely at least on the following two assumptions: First, a theory of rationality in game situations must assign one strategy to each player, and the

public announcement of the strategy combination should not prompt any player to deviate. Neither one of these assumptions seems compelling. Interestingly, this assumption that a rational theory for such situations should give a unique recommendation also occurs in John Nash's dissertation. He introduces two interpretations of Nash Equilibria. one of which is the rational prediction interpretation. He points out that a criterion of adequacy for a rational prediction is that the players should all be able to figure it out, that it should be unique and that nobody should act just out of conformity with the prediction. Nash, however, does not assume that such a prediction is possible for each game; he restricts attention so "solvable games", i.e., games which have a unique Nash Equilibrium or the exchangeability property of zero-sum games.

[20] Unless one interprets equilibria as equilibria *in beliefs*, as Aumann and Brandenburger (1995) suggest. That is, the probabilities would not express the randomization of an agent over his acts, but the joint beliefs of the other agents about what she is going to play. Equilibria would then be consistent systems of belief.

[21] Thucydides, *The Peloponnesian War*, Book VI, section 36.

[22] Dixit and Nalebuff (1991) provide a good example of how this argument is used. They envisage a situation where two newspaper managers have to set the price for their respective journals in view of how much the other will charge. "If he charges 1, I should charge 2. But he, knowing I am thinking in this way, will charge not 1, but his best response to my 2, namely 2.50. Then I should charge my best response, 2.75. But then he . . . ". After some pondering of this kind, Dixit and Nalebuff introduce their Rule 4: Having exhausted the simple avenues of looking for dominant strategies or ruling out dominated ones, the next thing to do is to look for an equilibrium of the game (pp. 74–7). A little reflection on this case, though, takes away much of its plausibility.

[23] Here is a more sophisticated version of the argument, following Skyrms (1990b): Suppose *n* agents have common knowledge of their prior probabilities and their update rules (which need to satisfy certain conditions for this argument to work). They embark on a process of changing their strategies into best responses to what the other agents are currently expected to do that converges to a Nash Equilibrium. That is, each starts with some probability distribution over his actions, which they modify in response to the initial probabilities of the others. In the next step, they modify their revised probabilities in response to the other agents' revised probabilities, etc. The assumptions on the update rules guarantee convergence. However, a little reflection shows that the discussion of the less sophisticated version applies point-by-point.

[24] And from there, Harsanyi and Selten (1988) go on to present a theory of equilibrium selection that picks a unique equilibrium for each game situation.

[25] Johansen (1982) derives Nash Equilibrium behavior from some rationality assumptions. One of them is a no-regret assumption, i.e., the assumption that a rational solution for game situations should assign actions to agents such that nobody will regret his choice at the end. So-called regret theories have received some attention in rational choice theory recently, cf. Sugden (1991).

[26] Cf. Aumann (1974), where the concept was introduced.

[27] An assumption of rationality at each state of the world implies, in virtue of this definition, that there is common knowledge of rationality.

[28] Cf. also Levi (1986), chapter 4.

[29] For a survey of the relevant issues, cf. Morris (1995).

[30] Cf. Jeffrey (1983, chap. 1), where the concept was originally introduced in the context of dealing with the Newcomb problem, and, about the connection between ratifiability

and Aumann's theorem, cf. Skyrms (1990a)). Readers who have never encountered this concept may wonder what its point is in the first place. Ratifiability makes a difference only in Newcomb-type situation, where the very fact that a decision has been made may change the considerations leading up to that decision. Only in such situations does this concept do any work.

[31] For another discussion of the connection between the concept of equilibrium and the concept of ratifiability, cf. Shin (1991).

[32] For helpful discussion or comments, I am indebted to Paul Benacerraf, Matthias Hild, Richard Jeffrey, Isaac Levi, and two anonymous referees for this journal.

REFERENCES

Arnauld, A. and P. Nicole: 1996, J. V. Buroker (ed.), *Logic or the Art of Thinking*, Cambridge University Press, Cambridge.

Aumann, R.: 1974, 'Subjectivity and Correlation in Randomized Strategies', *Journal of Mathematical Economics* **1**, 67–96.

Aumann, R.: 1985, 'What is Game Theory Trying to Accomplish?', in K. Arrow and S. Honkapojah (eds.), *Frontiers of Economics*, Blackwell, Oxford.

Aumann, R.: 1987a, 'Game Theory', in J. Eatwell et al. (eds), *The New Palgrave Dictionary of Economics*, Stockton, New York.

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

Aumann, R.: 1990, 'Nash Equilibria are not Self-Enforcing', in J. J. Gabszewicz, J. F. Richard, and L. A. Wolsey (eds), *Economic Decision Making: Games, Econometrics, and Optimization*, Elsevier, Amsterdam.

Aumann, R. and M. Maschler: 1972, 'Some Thoughts on the Minimax Principle', *Management Science* **18**(5).

Aumann, R. and A. Brandenburger: 1995, 'Epistemic Conditions for Nash Equilibrium', *Econometrica* **63**, 1161–80.

Bacharach, M.: 1987, 'A Theory of Rational Decision in Games', *Erkenntnis* **27**, 17–55.

Bernheim, D.: 1984, 'Rationalizable Strategic Behavior', *Econometrica* **52**, 1007–28.

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

Clausewitz, C. von: 1972, *Vom Kriege*, Duemmler Verlag, Bonn.

Dekel, E. and F. Gul: 1997, 'Rationality and Knowledge in Game Theory', in D. Kreps and K. Wallis (ed.), *Advances in Economics and Econometrics: Theory and Applications*, Chap. 5, Cambridge University Press, Cambridge.

Dixit, A. and B. Nalebuff: 1991, *Thinking Strategically*, W. Norton, New York.

Ellsberg, D.:1954, 'Theory of the Reluctant Duelist', in O. Younng (ed.), *Bargaining Formal Theories of Negotiation*, University of Illinois Press, Chicago.

Farrell, J.: 1988, 'Communication, Coordination and Nash Equilibrium', *Economic Letters* **27**, 209–14.

Fishburn, P.: 1981,'Subjective Expected Utility Theory: An Overview of Normative Theories', *Theory and Decision* **13**, 139–99.

Hampton, J.: 1994, 'The Failure of Expected Utility Theory as a Theory of Reason', *Economics and Philosophy* 195–242.

Harsanyi, J.: 1977, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press, Cambridge.

Harsanyi, J.: 1982, 'Subjective Probability and the Theory of Games: Comments on Kadane and Larkey's Paper', *Management Science* **28**(2).

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

Jacobsen, H. J.: 1996, 'On the Foundations of Nash Equilibrium', *Economics and Philosophy* **12**, 67–88.

Jeffrey, R.: 1983, *The Logic of Decision*, University of Chicago Press, Chicago.

Jeffrey, R.: 1992, *Probability and the Art of Judgement*, Cambridge University Press, Cambridge.

Johansen, L.: 1982, 'On the Status of the Nash Type of Non-cooperative Equilibrium in Economic Theory', *Scandinavian Journal of Economics* **84**, 421–41

Kadane, J. and P. Larkey: 1982, 'Subjective Probability and the Theory of Games', *Management Science* **28**(2).

Kalai, E. and E. Lehrer: 1993, 'Rational Learning Leads to Nash Equilibrium', *Econometrica* **61**, 1019–45..

Kreps, D.: 1989, 'Nash Equilibrium', in J. Eatwell et al. (eds), *The New Palgrave: Game Theory*, Norton, New York.

Kreps, D.: 1990, *Game Theory and Economic Modeling*, Clarendon Press, Oxford.

Levi, I.: 1986, *Hard Choices*, Cambridge University Press, Cambridge.

Levi, I.: 1997, 'Prediction, Deliberation, and Correlated Equilibrium', *The Covenant of Reason*, Cambridge University Press, Cambridge, pp. 102–17.

Lewis, D. K.: 1969, *Convention. A Philosophical Study*, Harvard University Press, Cambridge, MA.

Luce, D. and H. Raiffa: 1957, *Games and Decisions*, Wiley, New York.

Morris, S.: 1995, 'The Common Prior Assumption in Economic Theory', *Economics and Philosophy* **11**, 227–53.

Nash, J.: 1950, *Non-Cooperative Games*, Dissertation, Princeton University, Princeton.

Nozick, R.: 1990, *The Normative Theory of Individual Choice*, Garland, London

Resnik, M.: 1987, *Choices*, University of Minnesota Press, Minneapolis.

Rubinstein, A. and M. Osborne: 1994, *A Course in Game Theory*, MIT Press, Boston

Savage, L.: 1954, *The Foundations of Statistics*, Wiley, New York

Salonen, H.: 1992, 'An Axiomatic Analysis of the Nash Equilibrium Concept', *Theory and Decision* **33**, 177–89.

Schelling, T.: 1960, *The Strategy of Conflict*, Harvard University Press, Cambridge, MA.

Shin, H. S.: 1991, 'Two Notions of Ratifiability and Equilibrium in Games', in M. Bacharach and S. Hurley (eds), *Foundations of Decision Theory*, Blackwell, Oxford.

Skyrms, B.: 1990a, 'Ratifiability and the Logic of Decision', *Midwest Studies in Philosophy* **XV**.

Skyrms, B.: 1990b, *The Dynamics of Rational Deliberation*, Harvard University Press, Cambridge, MA.

Sugden, R.: 1993, 'Thinking as a Team: Towards an Explanation of Non-Selfish Behavior', *Social Philosophy and Policy* 69–88.

Sugden, R.: 1991, 'Rational Choice: A Survey of Contributions from Economics and Philosophy', *Economic Journal* **101**, 751–85.

Department of Philosophy
1879 Hall, Princeton University
NJ 08544, Princeton
U.S.A.
E-mail: mrisse@princeton.edu