Tilburg University

Equilibrium in strategic interaction

van Damme, E.E.C.; Weibull, J.W.

Published in:
The Scandinavian Journal of Economics

Publication date:
1995

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
van Damme, E. E. C., & Weibull, J. W. (1995). Equilibrium in strategic interaction: The contributions of John C. Harsanyi, John F. Nash and Reinhard Selten. *The Scandinavian Journal of Economics*, 97(1), 15-40.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Equilibrium in Strategic Interaction: The Contributions of John C. Harsanyi, John F. Nash and Reinhard Selten*

Eric van Damme

Tilburg University, The Netherlands

Jörgen W. Weibull

University of Stockholm, Sweden

I. Introduction

During the past 15 years or so a “game theoretic revolution of economics” has taken place. Microeconomic theory has been rewritten in the language of a game theory; see Kreps (1990b).¹ Publications in industrial organization differ substantially from work published in this area two decades ago. In macroeconomics, issues of time consistency, credibility and reputation have been analyzed by means of game theoretic tools. Trade theory has been cast in game theoretic terms. In financial economics, problems of market-micro structure are analyzed by means of non-cooperative games. Fundamental concepts and techniques developed by this year’s Nobel laureates have enabled this revolution to take place. As Kreps writes:

Nowadays one cannot find a field of economics (or of disciplines related to economics, such as finance, accounting, marketing, political science) in which understanding the concept of a Nash equilibrium is not nearly essential to the consumption of the recent literature. (1990a, p. 1)

Nash equilibrium embodies a fundamental idea of economics, that people act in accordance with their incentives; cf. Aumann (1987a). John Nash not only introduced this universal solution concept that later came to bear his name, he also showed that each finite game admits at least one solution. Nash also introduced the fundamental distinction between cooperative games, in which binding agreements are presumed, and non-cooperative

---

* We wish to thank the editors of this Journal, and Johan Stennek, for helpful comments on an earlier draft of this paper.

¹ Dates in brackets refer to publications by Harsanyi, Nash and Selten (see Bibliographies) and dates in parentheses to references listed at the end of this article.

² See Gibbons (1992) for a sample of applications and further references.

games, in which binding agreements are not feasible unless they are explicitly allowed by the rules of the game.

Nash's ideas were further developed by Harsanyi and by Selten. John Harsanyi showed that games with incomplete information could be remodeled as games with complete but imperfect information, thereby enabling analysis of this important class of games and providing a theoretical foundation for "the economics of information". Reinhard Selten was the first to refine the Nash equilibrium concept for analysis of dynamic strategic interactions. Such refinement is necessary since many equilibria involve non-credible threats and are unreasonable in economic terms. Selten's formalization of the requirement that only credible threats should be taken into account, the concept of subgame perfect equilibrium, is widely used in the industrial organization literature and has led to important insights there as well as in other fields of economics. Although Nash had already tried to reduce the number of equilibrium solutions to one, by imposing additional (stability) requirements, the first coherent general theory of equilibrium selection was constructed only much later in joint work by Harsanyi and Selten.

In this paper we review the three laureates' main contributions to non-cooperative game theory.³ Our emphasis is on the foundations and background of the equilibrium concepts that nowadays are routinely applied in economics. To provide a proper perspective, we start in Section II by reviewing the relationship between games and economic theory and by paraphrasing the desiderata for a theory of strategic interaction from the founding fathers of game theory, John von Neumann and Oskar Morgenstern. In Section III we introduce the Nash equilibrium concept, discuss Nash's existence proofs and his attempts to arrive at a unique solution, and briefly review his work on bargaining theory. Section IV is devoted to Selten's concepts of subgame perfection and perfect equilibrium. Section V considers Harsanyi's approach to games with incomplete information. In Section VI we briefly discuss the main elements of Harsanyi's and Selten's theory of equilibrium selection, in particular the concept of risk dominance and the tracing procedure.

The material in Sections II–VI rests on the interpretation of Nash equilibrium as a necessary requirement for prescriptions based on a theory of rational behavior in strategic interactions. According to this interpretation, the game is played exactly once, and the players are rational decision makers. An alternative interpretation, first formulated in Nash's unpublished doctoral thesis [1950a], is discussed in Section VII. In this interpretation the game is played over and over again by players who need

³ For further information about Nash we refer to Leonard (1994); about Harsanyi, see Selten [1992a, ii]; about Selten, see Selten [1993b]. For an excellent survey of the development of game theory, see Aumann (1987b).

not be rational and who are randomly drawn from large populations. This interpretation (until recently largely unknown) is relevant to the current literature on learning and social evolution in games. Section VIII concludes by looking back at von Neumann's and Morgenstern's desiderata for a theory of strategic interaction.

II. Games and Economic Behavior

Game theory is a mathematical method for modeling and analyzing strategic interaction. While economists such as Cournot (1838) devised methods for studying strategic market interaction more than 150 years ago, the first systematic attempts to analyze strategic interactions in a unified framework were provided in the fundamental book *Theory of Games and Economic Behavior* (1947) by John von Neumann and Oskar Morgenstern. In the first chapter of this book the authors claim that the theory of "games of strategy", initiated by von Neumann (1928), is the proper instrument with which to develop a theory of economic behavior. They do not just argue that there are similarities between games and economic situations, rather

We hope to establish satisfactorily (...) that the typical problems of economic behavior become strictly identical with the mathematical notions of suitable games of strategy. (p. 2)

After having discussed the common elements in some fundamental economic problems, they conclude:

For economic and social problems the games fulfil — or should fulfill — the same function which various geometrico-mathematical models have successfully performed in the physical sciences. (p. 32)

von Neumann and Morgenstern describe the goal which they want to achieve:

We wish to find the mathematically complete principles which define 'rational behavior' for the participants in a social economy and to derive from them the general characteristics of that behavior. (p. 31)

The general characteristics of rational behavior, i.e., the implications (outcomes) of a theory of rational behavior can in principle be found in the following way:

[We] use the conventional tools of logics, and in particular that of the *indirect proof*. This consists of imagining that we have a satisfactory theory of a certain desired type, trying to picture the consequences of this imaginary situation and then in drawing conclusions from this as to what the hypothetical theory must be like in detail. If this process is applied successfully, it may narrow down the possibilities for the hypothetical theory of the type in question to such an extent that only one possibility is left — i.e. that the theory is determined. (pp. 147–8)

von Neumann and Morgenstern also discuss the consequences of the theory being determined:

Let us now imagine that there exists a complete theory of the zero-sum two-person game which tells each player what to do, and which is absolutely convincing. If the players knew such a theory then each player would have to assume that his strategy has been "found out" by his opponent. The opponent knows the theory, and he knows that a player would be unwise not to follow it. Thus the hypothesis of the existence of a satisfactory theory legitimizes our investigation of the situation when a player's strategy is "found out" by his opponent. (p. 148)

In the formal development of their general theory, however, von Neumann and Morgenstern do not follow up on this idea. After having discussed the minimax solution of two-person zero-sum games in strategic form, they note that other types of games always involve a parallelism of interests between some players and, hence, are qualitatively different. In these games players have incentives to form coalitions and make side-payments (even outside the formal rules of the game). They argue that a theory cannot be formulated without such auxiliary concepts. Accordingly, their theory assumes that coalitions can be formed outside the game and that these are respected by the players. Nash [1951] introduces the terminology "cooperative games" for such games, but he also suggests that they can be fruitfully analyzed by "non-cooperative" techniques. Nash also notes that the above argument of von Neumann and Morgenstern retains its validity beyond the zero-sum case. He formalizes it and investigates its consequences.

III. Equilibrium Points in Non-Cooperative Games

Nash's [1951] paper opens by making clear the distinction between cooperative and non-cooperative games:

von Neumann and Morgenstern have developed a very fruitful theory of two-person zero-sum games in their book *Theory of Games and Economic Behavior*. This book also contains a theory of n-person games of a type which we would call cooperative. This theory is based on an analysis of the interrelationships of the various coalitions which can be formed by the players of the game.⁴

Our theory, in contradistinction, is based on the absence of coalitions in that it is assumed that each participant acts independently, without collaboration or communication with any others.

The notion of an equilibrium point is the basic ingredient in our theory. This notion yields a generalization of the concept of the solution of a two-person zero-sum game. (*op. cit.*, p. 286)

---

⁴ On p. 295, Nash writes that this theory assumes that "coalitions will be enforced by an umpire".

The non-cooperative idea is implicit in the formal definitions, which we now give. Nash restricts himself to finite n -person games in strategic, or normal, form. Such a game is a triplet $G = \langle I, A, u \rangle$ where $I = \{1, \dots, n\}$ is the finite set of players, $A = \times_{i \in I} A_i$ with A_i being the (finite) set of pure strategies of player i , and $u = (u_i)_{i \in I}$ with $u_i: A \rightarrow \mathbb{R}$ being the payoff function of this player. It is assumed that u_i is a von Neumann-Morgenstern utility function. A mixed strategy s_i of player i is defined as a probability distribution s_i on A_i and $s_{i\alpha}$ denotes the probability that $s_i \in S_i$ assigns to $\alpha \in A_i$. We write S_i for the set of all mixed strategies of player i , $S = \times_{i \in I} S_i$ for the set of mixed strategy profiles and $u_i(s)$ for player i 's expected payoff when s is played. A mixed strategy profile s is a (Nash) equilibrium point if

$$u_i(s) = \max_{s'_i \in S_i} u_i(s \setminus s'_i) \quad (\text{all } i \in I)$$

where $s \setminus s'_i$ denotes the strategy profile $(s_1, \dots, s_{i-1}, s'_i, s_{i+1}, \dots, s_n)$. Thus a strategy profile s is an equilibrium point if each player's mixed strategy is optimal against those of the others.⁵

Nash [1951] does not provide an interpretation of this solution concept. However, Nash explains its significance in the section "Motivation and Interpretation" of his (unpublished) doctoral thesis written at the Mathematics Department of Princeton University during 1949–50, under the guidance of Professor Albert Tucker. Actually, Nash provides two interpretations. We take up one here and the other, the less known "mass-action" interpretation, in Section VII. In his first interpretation Nash [1950a] is clearly inspired by von Neumann and Morgenstern:

We proceed by investigating the question: what would be a "rational" prediction of the behavior to be expected of rational playing the game in question? By using the principles that a rational prediction should be unique, that the players should be able to deduce and make use of it, and that such knowledge on the part of each player of what to expect the others to do should not lead him to act out of conformity with the prediction, one is led to the concept of a solution defined before (...). In this interpretation we need to assume that players know the full structure of the game in order to be able to deduce the prediction for themselves. It is quite strongly a rationalistic and idealizing interpretation. (p. 23)

⁵ Hence, each player has perfect foresight about the mixed strategies of the others, and each player optimizes given his expectations. The relation between perfect foresight and equilibrium had already been discussed informally by Morgenstern (1935) and Hayek (1937). After having discussed the Holmes/Moriarty story from Conan Doyle Morgenstern concludes that equilibrium is incompatible with perfect foresight (p. 344). Having learned about mixed strategies from von Neumann, he later abandoned this view; see von Neumann and Morgenstern (1947, p. 176). Hayek (1937) states that perfect foresight, i.e., that "every person's plan is based on the expectation of just those actions of other people which those other people intend to perform" (p. 42), is the defining characteristic of equilibrium. However, he does not explicitly assume optimization, nor does he formalize his idea.

The main mathematical result in Nash [1951] is a quite elegant existence proof based on Brouwer's fixed point theorem.⁶ For $s \in S$ and $\alpha \in A_i$ write $u_{i\alpha}^+(s) = \max\{0, u_i(s \setminus \alpha) - u_i(s)\}$.

Then player i has an incentive to deviate from s if $u_{i\alpha}^+(s) > 0$, and s is an equilibrium if and only if $\sum_{\beta \in A_i} u_{i\beta}^+(s) = 0$ for all $i \in I$. Define the map $F_i: S \rightarrow S_i$ by means of $F_i(s) = s'_i$ where

$$s'_{i\alpha} = \frac{s_{i\alpha} + u_{i\alpha}^+(s)}{1 + \sum_{\beta \in A_i} u_{i\beta}^+(s)} \quad (\text{all } \alpha).$$

Since F is a continuous mapping from the convex and compact set S into itself, it has a fixed point $\bar{s}$, hence $\bar{s}'_{i\alpha} = \bar{s}_{i\alpha}$ for all players i and pure strategies $\alpha \in A_i$. Consequently

$$\bar{s}_{i\alpha} \sum_{\beta \in A_i} u_{i\beta}^+(\bar{s}) = u_{i\alpha}^+(\bar{s}) \quad (\text{all } i, \alpha).$$

If the sum on the l.h.s. would be positive for some player i , then we would have

$$\bar{s}_{i\alpha} > 0 \Leftrightarrow u_i(\bar{s} \setminus \alpha) > u_i(\bar{s}).$$

for each of his pure strategies α . However, since $u_i(\bar{s}) = \sum_{\alpha} u_i(\bar{s} \setminus \alpha) \bar{s}_{i\alpha}$, this is impossible. Consequently, the sum must be zero and $\bar{s}$ is an equilibrium point.

The reader may recognize this argument from existence proofs for Walrasian equilibria: just replace the "excess payoffs" $u_i(s \setminus \alpha) - u_i(s)$ by excess demands. Indeed, the early developments of general equilibrium theory, and this method of proof in particular, were influenced by the work of Nash. Arrow and Debreu (1954) used Nash's method of proof and wrote "... Nash has formally introduced the notion of an equilibrium point for a game. The definition can easily be extended to an abstract economy" (p. 273).

Nash is aware of the fact that a game may have multiple equilibria and he introduces several concepts to deal with this multiplicity. In particular, a set E of equilibria is said to be *interchangeable* if, whenever $s, t \in E$, also $s \setminus t_i \in E$ for every $i \in I$. A set E is called a *subsolution* if it is a maximal interchangeable set, and a game is called *solvable* if its full set of equilibria is interchangeable. Nash notes that a two-person zero-sum game is always solvable in this sense. However, more general games need not be solvable.

⁶ Nash [1950b] contains a proof using the Kakutani fixed point theorem and he thanks David Gale for having suggested the use of this tool. In his thesis, written afterwards, Nash had given a different proof using Brouwer's theorem. In his 1951 paper, Nash writes that "the proof given here (i.e., the one that EvD and JWW reproduce here) is a considerable improvement on that earlier version".

Nash argues that a subsolution may be looked at as a set of mutually compatible equilibria, forming a coherent whole, and that in unsolvable games it sometimes happens that good heuristic reasons can be found for narrowing down the set of equilibria to those lying in a single subsolution.

One example of such a heuristic procedure is found in Nash's work on bargaining. Given a 2-person game $G = \langle I, A, u \rangle$, Nash [1953a] builds a non-cooperative bargaining game that involves four stages:

1. Each player i simultaneously commits to a mixed strategy in G which he will be forced to use in case they cannot come to an agreement in stage 2. This strategy t_i is player i 's threat.⁷
2. The players inform each other of their threats.
3. Players independently state *utility demands*, d_1 and d_2 , i.e., minimal utility levels that they require.
4. If the utility demands can be simultaneously satisfied, i.e., if there is a joint probability distribution—in modern language, a correlated strategy; cf. Aumann (1974)— μ over the set of action pairs A such that $u_i(\mu) \geq d_i$, then each player gets what he demanded; otherwise the threats must be executed, hence, in that case player i 's payoff is $u_i(t)$.

Nash first solves the stage 2 (sub)game induced by each pair t of threat strategies. He notes that typically the stage 2 game will have multiple equilibria: any pair of demands that is Pareto efficient in G and that Pareto dominates the threat point is an equilibrium. However, Nash notes that different equilibria have different stability properties and that by studying these relative stabilities a unique equilibrium can be selected.

Without loss of generality assume $u_i(t) = 0$ for $i = 1, 2$. Then, player i 's payoff in the subgame t is given by $u_i^t(d) = d_i \chi(d)$ where χ is the indicator function of the set B of feasible payoffs in G . Hence, $\chi(d)$ takes the value 1 if the demands $d = (d_1, d_2)$ are compatible and 0 otherwise.⁸

Note that the payoff function is discontinuous. Nash proposes approximating this discontinuous function with a continuous one. Let $h: \mathbb{R}^2 \rightarrow \mathbb{R}_+$ be a continuous function that takes the value 1 on B , and that tapers off very rapidly to zero as d moves away from B . Consider the game where player i 's payoff associated with demand vector d is $u_i^h(d) = d_i h(d)$. It is now easy to see that if d^* is a maximizer of $d_1 d_2 h(d)$ on $\mathbb{R}_+^2$, d^* is an equilibrium of this game. Nash remarks that when less and less smoothing

⁷ Nash writes that "it is essential for the success of the threat that A be *compelled* to carry out his threat T if B fails to comply. Otherwise it will have little meaning. For, in general, to execute the threat will not be something A would want to do, just for itself. (...) Thus we need a sort of umpire, who will enforce contracts or commitments." (p. 130)

⁸ Formally, $B = \{v \in \mathbb{R}_+^2 : v = u(\mu) \text{ for some probability measure } \mu \text{ on } S\}$, and $\chi(d) = 1$ if $d \in B$ and otherwise $\chi(d) = 0$.

is used (i.e., when h converges to the indicator function of B_i) any such maximizer must converge to the unique argmax of the function $d_1 d_2$ on B_i . Hence, Nash argues that the only necessary limit of equilibrium points of smoothed games is

$$d(t) = \max_{d \in B \mathbb{R}_+^2} d_1 d_2$$

and he proposed selecting this equilibrium as the solution of subgame t . Nash had derived this solution by means of the axiomatic method in an earlier paper [1950b].

Having solved each subgame in this way, Nash then replaces each subgame by its value and notes that the resulting stage 1 game is strictly competitive. Since all stage 2 outcomes are Pareto efficient, one player can gain only if the other loses. Consequently, all equilibria are interchangeable. Hence, the bargaining game is solvable.

The research program to underpin axiomatic models in cooperative game theory with non-cooperative models of an associated decision procedure has become known as the *Nash program*. According to this program, if binding agreements are possible the offers and counteroffers leading to them should be modelled as formal moves in a non-cooperative game:

The writer has developed a “dynamical” approach to the study of cooperative games based upon reduction to non-cooperative form. One proceeds by constructing a model of pre-play negotiation so that the steps of negotiation become moves in a larger non-cooperative game (...) describing the situation. This larger game is then treated in terms of the theory in this paper (...) and if values are obtained they are taken as the values of the cooperative games. Thus the problem of analyzing a cooperative game becomes the problem of obtaining a suitable, and convincing, non-cooperative model for the negotiation. [Nash, 1951, p. 295]

Early contributions to this program are Harsanyi [1974] and Selten [1981]. Harsanyi’s model may be viewed as a dynamic n -person version of Nash’s four stage bargaining model and it leads to a better understanding of the so-called stable sets introduced by von Neumann and Morgenstern. In Selten’s model, players successively make proposals for coalitions and payoff divisions until an offer is accepted by all members of a coalition. In the 2-player case, Selten’s game form is similar to that of Rubinstein (1982). This latter paper, however, allows for a richer time preference structure and establishes uniqueness of the (subgame perfect) equilibrium outcome under general conditions of impatience. (Another precursor of Rubinstein (1982) is Ståhl (1972) where a finite alternating offers procedure is analyzed.) For the special case where the two players have the same time preference, Rubinstein’s alternating-offers bargaining model yields Nash’s axiomatic bargaining solution in the limit when negotiation

rounds become virtually instantaneous. The models of Harsanyi, Selten, Ståhl and Rubinstein all are extensive form games and they do not just use Nash equilibrium, they use Selten's [1965a,b] concept of subgame perfect equilibrium, the topic of the next section.

IV. Dynamic Interactions and Perfect Equilibria

Nash [1950a, 1950b, 1951] proved the existence of equilibria for games in strategic form. Since each game in extensive form can be represented in strategic form, Nash's existence theorem implies that any finite extensive form game has at least one equilibrium.

von Neumann and Morgenstern argued that a rational player should have no difficulty in making up his mind in advance for all possible contingencies. Accordingly, they claimed that the extensive and strategic forms are fully equivalent so that researchers can choose to work with whichever form is most convenient. They also argued that the strategic form, being analytically simpler, is more suited to deriving general theorems, while the extensive form is more suited to analyzing specific games. Indeed, in his analysis of a simplified poker game, Nash worked directly with the latter, using what Kuhn (1950, 1953) came to call behavior strategies; see Nash and Shapley [1950].⁹

This equivalence of the two game forms was, however, challenged by Reinhard Selten [1965a,b and 1975a]. He argues that players in an extensive game have no possibility to commit themselves to their strategies in advance and that the strategic-form analysis to a certain extent assumes such commitment.¹⁰ Selten [1965a] also notes that the impossibility of self-commitment¹⁰ implies that certain Nash equilibria cannot be viewed as "reasonable" solutions:

"Wenn man von der Annahme des völligen Fehlens jeder Selbstbindungskraft ausgeht, kann nicht jeder Gleichgewichtspunkt als eine vernünftige Lösung angesehen werden." (p. 307)

Selten gives the example in Figure 1 to illustrate why this is the case. The game is played as follows. First player 1 decides between l_1 and r_1 . If he chooses r_1 the game ends and each player receives the payoff 2. If l_1 is

---

⁹ Nash and Shapley note that the restriction to behavior strategies would not be legitimate if each player's information were not monotone increasing. Conversely, Kuhn (1953) shows that the procedure is legitimate if information is monotonically increasing (i.e., if the game has perfect recall).

¹⁰ Selten writes that the importance of self-commitment was first stressed in Schelling (1960). Schelling mainly discusses the advantages of being able to self-commit and investigates the possibilities for obtaining such commitment; in contrast, Selten investigates the consequences of the absence of self-commitment.

Fig. 1.

chosen, player 2 has to move. If he chooses l_2 player 1 receives 3 and player 2 receives 1; if he chooses r_2 each player has payoff zero. This game has two equilibria in pure strategies, viz. $L = (l_1, l_2)$ and $R = (r_1, r_2)$. Selten writes that an interpretation of equilibrium R requires that player 2 threatens before the start of the game that he will play r_2 in order to induce player 1 to play r_1 . Such a threat, however, is empty when player 2 cannot commit himself to actually carrying it out. Since the game is assumed non-cooperative, such self-commitment is impossible, hence, the threat is vacuous. Facing the *fait accompli* that player 1 has chosen l_1 , player 2 responds optimally by playing l_2 . Consequently, it is optimal for player 1 to play l_1 and only the equilibrium L is “reasonable”.

The above example shows that we need a more refined solution concept. Selten argues that when players cannot bind themselves, behavior in a subgame (i.e., a part of the game that constitutes a game in itself) should constitute an equilibrium of the subgame. Otherwise, if there would exist a subgame in which the solution does not constitute an equilibrium, then at least one player would have an incentive to deviate from that solution if the subgame were actually reached, and the solution would not be self-enforcing. Selten [1965a, b] calls such equilibria perfect, but in a later paper [1975a] he renames them *subgame perfect equilibria*. In Figure 1, only the equilibrium L is subgame perfect.

Selten [1965a, b] uses the subgame perfectness concept to solve an oligopoly game with demand inertia. The demand of a firm in a period does not just depend on the prices set by the firms in this period but also depends on the market shares from the previous period. In the finite horizon case, Selten determines a subgame perfect equilibrium by means of Bellman's (1957) dynamic programming technique, while he solves the infinite horizon problem by investigating the limit of the finite horizon case.

Of course, Selten was not the first to use a backward induction procedure in games. As early as 1912, E. Zermelo had used this procedure to argue that chess is completely determined, i.e., that either

white possesses a strategy that guarantees winning, or that black possesses a strategy that guarantees winning, or that each side possesses a strategy that guarantees at least a draw no matter what the other side does. More generally, Kuhn (1953) had shown that each finite extensive form game with perfect information has an equilibrium in pure strategies. Kuhn introduced the notion of a subgame and showed that each equilibrium of a subgame and an equilibrium of the associated truncated game (where the subgame is replaced by the selected equilibrium payoffs) can be combined to an equilibrium of the overall game.

In view of current concerns about the validity of backward induction, it is interesting to note that the founding fathers of game theory had raised an objection to this induction procedure. According to von Neumann and Morgenstern (1947): "This approach (...) presupposes 'rational' behavior of all players; i.e. player 1's strategy is based on the assumption that player 2's strategy is optimal and vice versa" (p. 127). The authors had already argued before that the hypothesis that others are always "rational" must be avoided. The argument is that a satisfactory and exhaustive theory of rational behavior can only be developed in a context that allows for irrational behavior. To know whether an action should be termed irrational or not, one has to evaluate its consequences, i.e., one has to know how a rational player would react to it. von Neumann and Morgenstern write:

"Imagine that we have discovered a set of rules for all participants — to be termed as "optimal" or "rational" — each of which is indeed optimal provided that the other participants conform. Then the question remains as to what will happen if some of the participants do not conform. If that should turn out to be advantageous to them — and, quite particularly, disadvantageous to the conformist — then the above "solution" would seem very questionable. (...) if the superiority of "rational behavior" over any other kind is to be established, then its description must include rules of conduct for all conceivable situations — including those where "the others" behaved irrationally, in the sense of the standards that the theory will set for them. (p. 32)¹¹

Selten is aware of the conceptual difficulties with the assumption of persistent rationality. At the same time, however, he realizes that the subgame perfectness concept is not restrictive enough since it allows some "unreasonable" equilibria to survive. Selten's ingenious suggestion for resolving both difficulties at once is to view perfect rationality as a limiting case of incomplete rationality by introducing the possibility that players

---

¹¹ Note the tension between this requirement to allow for (others') irrationality and the earlier claim that the strategic and extensive forms are equivalent — a claim that was based on the presumption that all players are fully rational.

may make slight mistakes when implementing their desired actions.¹² He writes [1975a]:

There cannot be any mistakes if the players are absolutely rational. Nevertheless, a satisfactory interpretation of equilibrium points in extensive games seems to require that the possibility of mistakes is not completely excluded. This can be achieved by a point of view which looks at complete rationality as a limiting case of incomplete rationality. (p. 35)

Formally, Selten assumes that at each information set (decision point) of each player there is a small, but positive probability that rationality breaks down. Whenever rationality breaks down, the choice of the player is governed by some random mechanism that selects each choice with a strictly positive probability. All probabilities are assumed to be known and independent of each other. Rational players will have to take the possibility of irrational play into account when making their decisions and they will represent their decision problem by a *perturbed game* that takes the mistakes explicitly into account. A slight modification of Nash's existence theorem implies that each such a perturbed game has at least one equilibrium. Selten defines an equilibrium to be (*trembling hand*) *perfect* if it is the limit of some sequence of Nash equilibria in perturbed games in which the error probabilities tend to zero, hence, the equilibrium can be rationalized by *some* form of slight irrationality. Since strategy spaces are compact, existence of perfect equilibria is easily established. Note that players are still "rational" in the sense that they know, and take account of, the mistake probabilities in their own and others' behavior; this is the rationale for considering Nash equilibria in the perturbed games.

The following numerical example may illustrate this procedure as well as the difference between perfect and subgame perfect equilibria. The game starts with a move of player 1. If this player chooses o_1 the game ends and both players receive 2. Otherwise player 2 will have to move and

Fig. 2.

¹² Selten attributes this idea by basing the definition of perfect equilibria on a model of slight mistakes to John Harsanyi.

this player then does not know whether player 1 has chosen l_1 or r_1 . This is indicated in Figure 2 by the dashed line connecting the two decision points of player 2. Hence, this player cannot condition his choice on whether player 1 has chosen l_1 or r_1 . The payoffs to the two players are given at the end points, the top number always being the payoff to player 1.

This game does not have any subgames. Hence, each Nash equilibrium is subgame perfect. In particular, this holds for (o_1, r_2) when, say, $x = 0$. Nevertheless, this equilibrium is not credible. If player 2 has to move, he knows that player 1 has chosen l_1 or r_1 . In both cases l_2 yields a strictly higher payoff than r_2 . Hence, player 2's optimal choice is l_2 and thus player 1 has an incentive to play l_1 . If player 1 with a small probability makes mistakes, then both l_1 and r_1 occur with positive probability and player 2 is forced to choose l_2 . The unique equilibrium of a slightly perturbed game is thus (l_1, l_2) . Hence, only (l_1, l_2) is a perfect equilibrium of the original game.

It is worthwhile to remark that for equilibrium to be termed perfect it is sufficient that it can be rationalized by *some* form of irrationality. For example, in the game from Figure 2 with $x = 1$, (o_1, r_2) is a perfect equilibrium since it is optimal for player 2 to choose r_2 if he believes that the mistake r_1 is sufficiently more likely than the mistake l_1 . In this game, (l_1, l_2) is also a perfect equilibrium and it is "more stable" than (o_1, r_2) since, once players expect (l_1, l_2) to be played, they will indeed play it, no matter what slight form of irrationality they expect. In general, there need not be an equilibrium that is stable with respect to *all* slight trembles. However, Kohlberg and Mertens (1986) and Mertens (1989) have shown that there always exists a minimal set of Nash equilibria with this property. Although Selten is of the opinion that requiring this strong form of stability is too much to ask, his work, as well as that of Harsanyi, has been an important source of inspiration for it; cf. Mertens (1989, Sect. 1.4).

The above example has shown that different perfect equilibria may have different degrees of plausibility and that such plausibility is naturally discussed in terms of the beliefs of players that sustain the equilibrium. This Bayesian language using beliefs originates in Kreps and Wilson (1982), where a slight modification of Selten's perfect equilibrium, the *sequential equilibrium* concept, is introduced. Kreps and Wilson build up their solution concept by combining a criterion of sequential rationality with a criterion of consistency in players' beliefs, which are formed in a Bayesian fashion. Indeed, they combine elements from Selten's and Harsanyi's work: "We have depended greatly and borrowed unconstrainedly from the insightful work of John Harsanyi and Reinhard Selten." (1982, fn. 1). Because the language of beliefs is more natural to economists than the language using mistakes, and since sequential equilibria are somewhat easier to work with than perfect equilibria, the sequential equilibrium concept has been used more frequently in applications.

V. Incomplete Information and Bayesian Equilibria

The discussion in the previous sections assumed each player to have full knowledge of the game; not only is he assumed to know the rules of the game in full detail, but also the payoff functions of the other players. In the early phase, game theoretic models were restricted to such situations of complete information. von Neumann and Morgenstern (1947) wrote:

Subsequent discussions will show that we cannot avoid the assumption that all subjects of the economy under consideration are completely informed about the physical characteristics of the situation in which they operate and are able to perform all statistical, mathematical, etc. operations which this knowledge makes possible. The nature and importance of this assumption has been given extensive attention in the literature and the subject is probably very far from being exhausted. We propose not to enter upon it. The question is too vast and too difficult and we believe that it is best to 'divide difficulties'. (p. 30)

The restrictiveness of this assumption was not unnoticed. For example, Luce and Raiffa (1957) wrote that "this is a serious idealization which only rarely is met in actual situations" (p.49) and that, rather than the rationality assumption, "the knowledge assumption appears to be the real source of unreality in the model" (p. 55). Although the importance of "incomplete" information was realized, no adequate tools for handling such situations were available, since they seemed to give rise to an infinite regress in reciprocal expectations.

Consider, for example, the negotiations between a buyer and seller concerning the sale of an asset. Assume that the buyer is privately informed about his willingness to pay and that also the seller has private information concerning his own reservation value. Consider the most simple bargaining game: the buyer and seller simultaneously quote prices and a transaction takes place — at the average of the two prices — if and only if the bid exceeds the ask. To optimally determine his ask price, the seller has to estimate the bid price and since the latter depends on the buyer's willingness to pay, the seller must form beliefs about what the willingness to pay is. Let us call these beliefs the seller's *first-order* expectations. Similarly, the buyer has first-order expectations about the seller's reservation value. Since the buyer's bid depends on the buyer's first-order expectations, the seller has to form beliefs about what the buyer's first-order beliefs will be. These beliefs may be called the *second-order* expectations. Since each player's bid will depend on this player's second-order expectation, the opponent will have to have *third-order* expectations, etc. *ad infinitum*.

Motivated by problems of arms control, Harsanyi [1967, 1968a,b] proposes an alternative, Bayesian, approach to the analysis of incomplete

information in order to cut this Gordian knot.¹³ The Bayesian approach requires that all uncertainty be made explicit and be quantified. To achieve this, Harsanyi postulates that each player may be of several possible *types*, where a type specifies *all* information that a player has about the game. In particular, a player's type specifies a probability distribution over the set of possible types of the other players in the game. One may think of a player's type as this player's state of mind and of a vector of types (one type for each player) as a state of the world. Indeed, knowing a player's type one can, in principle, compute this player's infinite hierarchy of beliefs. However, the main advantage of the "type construction" is that there is no need to study this hierarchy. Formally, Harsanyi defines a *Bayesian game* as a tuple $\langle I, S, T, u, p \rangle$, where I is the player set, S_i is the strategy set for player i , $T = \times_i T_i$ where T_i is the set of types of player i , $u = (u_i)_i$ with $u_i: S \times T \rightarrow \mathbb{R}$ being the payoff function of this player, and $p = (p_i)_i$ with $p_i(t_{-i}; t_i)$ being the posterior belief of type t_i of player i concerning the types $t_{-i} = (t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_n)$ of the other players. A *Bayesian (Nash) equilibrium* specifies a strategy for each type such that each type maximizes its conditional expected payoff taking the other types' strategies as given.

A further simplification can be obtained if the players' beliefs about each others' types are derivable from a common prior distribution on the product space of types in the game, i.e., if there exists a probability distribution q on T such that $q(t_{-i}|t_i) = p_i(t_{-i}; t_i)$ for all t . In this so-called *consistent* case, there exists a simple game of complete, but imperfect information that is equivalent to the original game of incomplete information.¹⁴ This *C-game* starts with a chance move that informs each player of his type, thereafter players choose actions. Harsanyi argues that it should always be possible to trace differences in rational players' beliefs back to differences in players' information, i.e., that a carefully constructed model should be consistent, a position that is referred to as the *Harsanyi doctrine*. Most applications of the theory are indeed restricted to this consistent case.

As an example, consider once more the negotiations between a buyer and a seller. Suppose that the *only* piece of private information that each player has is the value he attaches to the object. A state of the world then specifies a value for each player and, in the consistent case, players have a common prior over the set of possible value pairs. A strategy for the seller (s) is then a map f that assigns an asking price $p = f(v_s)$ to each value v_s that

¹³ Harsanyi's work was instrumental to a joint research project for the US Arms Control and Disarmament Agency, also involving, among others, R. Aumann, M. Maschler, R. Selten and R. Stearns; see *Mathematica* (1966, 1967, 1968).

¹⁴ Also the inconsistent case can be reduced to a game of imperfect information by introducing appropriate lotteries at the end of the game, an idea that Harsanyi attributes to Selten.

the seller might have. Similarly, a strategy for the buyer (b) specifies a bid $q = g(v_b)$ for each $v_b \in [0, 1]$. A strategy pair (f, g) is a *Bayesian (Nash) equilibrium* if and only if each player maximizes his expected revenue for each realization of his own value. In equilibrium, the buyer “shades” his offer below his true reservation price and the seller’s strategy calls for a “mark up”, with the consequence that not all possible gains from trade are exploited; see Chatterjee and Samuelson (1983). In fact, it can be shown that, no matter how one constructs the rules of the trading game, the gains from trade will not be fully exploited; see Myerson and Satterthwaite (1983). In a corresponding game of complete information, such a loss does not occur in equilibrium. Harsanyi’s method thus allows us to identify welfare losses due to the incompleteness of information in situations of strategic interaction.

Harsanyi’s work may be viewed as the foundation for nearly all economic analysis involving differential information and today this area is one of the most active in economic theory. The modeling and analysis of the above example has become a standard tool of the trade and nowadays it may be hard to understand why incomplete information games posed such great difficulties before Harsanyi’s conceptual innovation. Of course, this testifies to the greatness of Harsanyi’s breakthrough. Actually, this example might fail to bring out the richness and depth of Harsanyi’s construction. For example, the situation in which each player does not precisely know the distribution from which the opponent’s value is drawn can also be handled by the same methods; one just has to enlarge the type space. In general, the less information the players have in common, the larger the types spaces in the Bayesian model have to be. Mertens and Zamir (1985) have shown that no matter what private information the players are supposed to have, there is always a mathematically well-defined type space that is large enough to represent this information. Hence, any incomplete information game can be represented by a Bayesian game — Harsanyi’s construction is completely general.

It should be noted that Vickrey (1961) had already analyzed (first-order) incomplete information in the special case of auctions. Among others, Vickrey analyzed the “Dutch auction” (of a single indivisible object) in which the auctioneer announces prices in descending sequence, with the first bid leading to an immediate transaction at that price. Vickrey noted that this situation can be viewed as a game and he argues that the appropriate solution concept is that of Nash equilibrium. By computing the equilibria explicitly, Vickrey establishes the revenue equivalence of the Dutch and English auctions for the symmetric case (all values of all players drawn from the same distribution). He also shows that the Dutch auction may produce inefficient outcomes in the asymmetric case, implying that revenue equivalence fails in this case. Vickrey also notes that a bidder

whose value is drawn from a finite set may have to use a randomized strategy in equilibrium, and that there is no need for randomization for a bidder whose values are drawn from a continuous distribution, i.e., there is no need to add strategic uncertainty if the uncertainty in the environment is sufficiently rich.

This last result of Vickrey gives a hint toward another important result of Harsanyi [1973a] concerning the purification of Nash equilibria. To establish his existence theorem, Nash, following von Neumann and Morgenstern, was forced to introduce mixed strategies, i.e., probability distributions over pure strategies. The interpretation of mixed strategies as deliberate acts of active randomization has, however, always met with difficulties. Furthermore, mixed equilibria are not strict and, in order to maximize his payoff, a player actually need not randomize: he could just as well play any pure strategy in support of the mixture. The only reason a player randomizes is to keep the other players in equilibrium. Consequently, mixed strategy equilibria do not seem to have a firm basis in Bayesian rationality. Such a basis was provided in Harsanyi [1973a] in which a Bayesian interpretation of mixed strategies is derived. A mixed strategy of player i is viewed as a representation of the ignorance that the opponents of i have about what i is actually going to do. Even though i may follow a deterministic rule, the opponents may not be able to predict exactly what i will do since this action may depend on information that the opponents can only assess probabilistically. Harsanyi argues that each player will always have at least a tiny bit of private information about his own payoffs and he modifies the game accordingly. Such a slightly perturbed version (of a strategic form game) always has an equilibrium in pure strategies and all regular equilibria of the original game may be interpreted as the limiting beliefs associated with these pure equilibria of the perturbed games. Hence, mixed strategies express ignorance; they should not necessarily be viewed as acts of conscious randomization.

VI. Equilibrium Selection

The “theory” rationale underlying the Nash equilibrium concept, i.e., the argument that a non-self-defeating theory of rationality has to prescribe a Nash equilibrium in every game, relies essentially on the assumption that the indirect method will indeed eliminate all but one strategy profile from the game (see quotations in Sections II and III). However, since in general a game may admit more than one Nash equilibrium, this rationale is incomplete, unless it is possible to select one unique equilibrium from the set of all Nash equilibria. Such a selection theory was developed by Harsanyi and Selten in their 1988 book.

Harsanyi and Selten write that “the basic task of game theory is to tell us what strategies rational players will follow and what expectations they can rationally entertain about each other. Of course these problems must be solved jointly since for a player’s strategy to be a rational strategy, it must be a best reply to his rational expectations about the other players’ strategies” (p. 342). Harsanyi and Selten insist that these rational expectations are strictly endogenous, in the sense of being based solely on the structure of the game (and being invariant with respect to the labeling of players and pure strategies). Consequently, the solution of a game should be independent of the context in which the game arises. In particular, the solution of a symmetric game should be symmetric, and the solution of a subgame should depend only on the subgame itself. This property of “subgame consistency” is quite clearly a strong form of the “persistent rationality” requirement and it allows the analyst to find the solution of an extensive form game by backward induction.

In the remainder of this section we restrict ourselves to games given in strategic form. For such games, Harsanyi and Selten propose a process of elimination and substitution — based on dominance relationships between equilibria — that reduces each (finite) candidate set to a unique solution. The core of the theory consists of a procedure that selects a unique equilibrium in each situation where there are two equilibrium candidates. A simple example of such a situation is the game given in Figure 3. The game is symmetric and the players have common interests. There are two equilibria in pure strategies: (α, α) and $(\bar{\alpha}, \bar{\alpha})$. The first Pareto dominates the second, but playing α is also more risky: if I play α , but my opponent plays $\bar{\alpha}$, my payoff is only zero. In contrast, if I play $\bar{\alpha}$ and my opponent α , my payoff is 3. In such a case, $(\bar{\alpha}, \bar{\alpha})$ is said to risk dominate (α, α) .

We now define risk dominance. Let G be a 2×2 game in which each player has the action set $\{\alpha, \bar{\alpha}\}$ and in which both (α, α) and $(\bar{\alpha}, \bar{\alpha})$ are strict equilibria, i.e., each player loses by deviating. Write $d_i(\alpha)$ for the loss that i incurs if he unilaterally deviates from (α, α) : $d_1(\alpha) = u_1(\alpha, \alpha) - u_1(\bar{\alpha}, \alpha)$. Identify a mixed strategy of player j by the probability

	α	$\bar{\alpha}$
α	4,4	0,3
$\bar{\alpha}$	3,0	2,2

Fig. 3.

a_j that this player assigns to α and write $\bar{a}_j = 1 - a_j$. The risk that player i is willing to take at (α, α) can be measured by the critical probability a_j^* that makes player i indifferent between playing α and $\bar{\alpha}$: the smaller a_j^* , the safer the equilibrium (α, α) for player i . The stability of the equilibrium (α, α) can be measured by the area of its stability region, i.e., where α is a best response for each player. The equilibrium (α, α) is said to *risk dominate* $(\bar{\alpha}, \bar{\alpha})$ if it has a larger stability region, i.e., if $\bar{a}_1^* \bar{a}_2^* > a_1^* a_2^*$, or equivalently, if $a_1^* + a_2^* < 1$. One can easily check that (α, α) risk-dominates $(\bar{\alpha}, \bar{\alpha})$ if and only if

$$d_1(\alpha) d_2(\alpha) > d_1(\bar{\alpha}) d_2(\bar{\alpha}),$$

i.e., if the product of the deviation losses at (α, α) is larger than that at $(\bar{\alpha}, \bar{\alpha})$. Note that this latter criterion is very much like the Nash product criterion that we encountered in Section III. Indeed the desire to have a solution with this “Nash product property” was an important guiding principle for Harsanyi and Selten in developing their theory. For 2×2 games, Harsanyi and Selten characterize this risk-dominance relation by a compelling set of axioms.

Returning to the game of Figure 3, we see that indeed (α, α) Pareto dominates $(\bar{\alpha}, \bar{\alpha})$, but that the risk-dominance comparison favors $(\bar{\alpha}, \bar{\alpha})$. In such cases of conflict between two selection criteria, the theory gives precedence to payoff dominance. Risk dominance is considered to be important only in those cases where players initially are uncertain about which equilibrium to choose. When one equilibrium Pareto dominates another, such uncertainty is assumed to be absent. Hence, the theory incorporates elements of both *collective* rationality (payoff dominance) as well as *individual* rationality (risk dominance). In the postscript of their book, Harsanyi and Selten motivate extensively why they made this choice and they discuss whether or not a theory that is based solely on risk dominance would be desirable. Note that a theory that incorporates payoff dominance criteria cannot be ordinal, i.e., the solution cannot just depend on the best-reply structure of the game. For example, the game from Figure 3 is “best reply equivalent” to the coordination game in which the payoff to (α, α) is $(1, 1)$ and that to $(\bar{\alpha}, \bar{\alpha})$ is $(2, 2)$ (off-diagonal payoffs are zero), but in this game $(\bar{\alpha}, \bar{\alpha})$ is payoff dominant. In contrast, risk dominance is an ordinal concept.

The conflict illustrated by the game in Figure 3 is just one instance where there is an irreconcilable conflict between two desirable properties. Chapter 3 of Harsanyi and Selten’s 1988 book shows just how quickly various desirable properties are conflicting; hence, essential choices have to be made and their theory is not the only possible or reasonable one. Indeed the authors themselves have subsequently suggested alternative

theories; see Harsanyi [1993c] and Selten [1994]. In other words, Harsanyi and Selten have shown that a coherent theory of rational behavior in games can be constructed. However, multiple theories are possible.

For games larger than 2×2 , risk dominance is defined by means of the tracing procedure that was originally developed in Harsanyi [1975c]. The tracing procedure is a model of a reasoning process that transforms players' prior beliefs into an equilibrium of the game. Formally, it is a map T that takes each mixed strategy profile s in a game G to an equilibrium $T(s)$ of the game. If the initial beliefs of rational players are described by s , these players will eventually reason themselves to $T(s)$ and they will coordinate on this equilibrium. (Mathematically, T is constructed from a homotopy that converts an equilibrium from a trivial game into the distinguished equilibrium $T(s)$ of the game under study. The homotopy parameter $t \in [0, 1]$ describes the degree of trust that the players put in the solution at "time" t .) Obviously, the outcome of the tracing procedure generally depends on the prior s , hence, a theory of rational behavior has to specify how the prior should be constructed.

Harsanyi and Selten propose the following construction based on the principle of insufficient reason. Assume that s and $\bar{s}$ are the solution candidates. Player i , being Bayesian, will attach a probability z_i to the solution being s and the complementary probability $1 - z_i$ to the solution being $\bar{s}$. Consequently he will be tempted to play the best response $b_i(z_i)$ against $z_i s + (1 - z_i) \bar{s}$. An opponent j can mimic this reasoning, but he will not be able to determine player i 's beliefs z_i exactly. Harsanyi and Selten assume that j considers z_i to be uniformly distributed on $[0, 1]$ and, consequently, that j expects i to play $\sigma_i = \int_0^1 b_i(z_i) dz_i$. The mixed strategy combination σ defined in this way should be taken as the starting point of the tracing procedure: s risk dominates $\bar{s}$ if and only if $T(\sigma) = s$.

As an important by-product of the research that eventually led to the tracing procedure, Harsanyi [1973b] obtained a proof that generic strategic form games have an odd number of equilibria. The proof was inspired by Debreu's (1970) proof that generic exchange economies have an odd number of Walrasian equilibria. Hence, we have come full circle: Nash's proof of the existence of strategic equilibria inspired a proof to establish existence of Walrasian equilibria. The demonstration of local uniqueness and stability of Walrasian equilibria by Debreu led to a similar result for strategic equilibria.

VII. Bayesian Rationality or Boundedly Rational Adaptation?

In the previous sections attention has been confined to what might be called "prescriptive" game theory: the theory deduces what outcomes are consistent with it being common knowledge that players are rational in the

Bayesian sense. Of course, if one wants to apply this theory, one is faced with the question of the descriptive relevance of a theory that is based on such strong rationality assumptions. Harsanyi [1977a, Sect. 2.1] contains a sophisticated discussion concerning this question; the following quote illustrates his optimistic attitude:

To be sure, it has been a matter of continual amazement among philosophers, and often even among social scientists (at least outside the economics profession), how any theory of rational behavior can ever be successful in explaining or predicting real life behavior. Yet it is hardly open to doubt that in actual fact such theories have been remarkably successful in economics and more recently in several other social sciences, particularly political science, international relations, organization theory, and some areas of sociology. (p. 16)

As a matter of fact, Harsanyi takes the principle of Bayesian rationality as a point of departure in most of his work, not only in game theory, but also in his research on moral decision making (leading there to a justification of utilitarianism) as well as in philosophy. We refer to Selten [1992a, ii] for an appreciation of this work of Harsanyi.

Nash had already pointed out, however, that strong rationality assumptions may not be necessary to justify his equilibrium concept. In his unpublished thesis [1950a], he introduces the “*mass action*” interpretation, which he argues is appropriate in many instances: “There are situations in economics or international politics in which, effectively, a group of interests are involved in a non-cooperative game without being aware of it; the non-awareness helping to make the situation truly non-cooperative.” He continues:

In this interpretation (...): It is unnecessary to assume that the participants have full knowledge of the total structure of the game, or the ability and inclination to go through any complex reasoning processes. But the participants are supposed to accumulate empirical information on the relative advantages of the various pure strategies at their disposal. (...) We assume that there is a population (...) of participants (...) and that there is a stable average frequency with which pure strategy is employed by the “average member” of the appropriate population. (p. 21)

Nash concludes that “this ‘mass-action’ interpretation leads to the conclusion that the mixed strategies representing the average behavior in each of the populations form an equilibrium point.” Of course, this interpretation of Nash equilibrium relies essentially on the assumption that there is a stable frequency and Nash himself did not indicate under which conditions such stability could be expected. For the special case of the Brown-Robinson process of fictitious play, an example of Shapley (1964) indicated that stability was non-generic.

Until recently, the mass-action interpretation of Nash [1950a] seems to have been largely unknown. However, evolutionary biologists independently introduced similar interpretations in the 1970s. In this branch of

game theory, called evolutionary game theory, there is no conscious choice at all. Individuals are biologically or socially conditioned or “programmed” to play certain strategies, and more successful strategies reproduce faster in the population than others. If the population reaches a dynamically stable state (distribution over strategies), all strategies present must be equally successful, and, due to stability, no better pure strategy is available. Hence, the population state must constitute a Nash equilibrium. This evolutionary approach was first proposed in a static form in Maynard Smith and Price (1973), and in an explicitly dynamic form in Taylor and Jonker (1978). A key notion is that of an *evolutionarily stable strategy* (ESS). A mixed strategy s^* in a symmetric two-player game is evolutionary stable if any alternative strategy $s \neq s^*$ that enters a population of individuals playing s^* in a sufficiently small fraction earns a lower post-entry payoff than the incumbent strategy s^* . It is easily shown that this condition is equivalent to (s^*, s^*) being a Nash equilibrium of the game with the additional property that s^* does better against any alternative best reply s to s^* than s does against itself. Hence, the ESS concept is a refinement of the Nash equilibrium concept. To increase the domain of applicability of this concept, it has been extended to asymmetric games as well as to extensive form games. Key contributions in this area were made in Selten [1980b, 1983c, 1988b]. The first applications to economics are now beginning to appear.

Today, the properties of various alternative dynamic adaptation and learning processes are being vigorously investigated. Along with evolutionary game theory, this is perhaps the most actively researched area in game theory. For recent contributions we refer to the proceedings of the Nobel symposium on Rationality and Equilibrium in Strategic Interaction (*Games and Economic Behavior*, vol. 8, no. 1, Jan. 1995). However, let us note that some earlier mentioned concepts such as risk dominance have also been found to play an important role in this interpretation. For instance, it has been shown that certain stochastic learning and adaptation processes in the stag hunt game from Figure 3 will eventually lead to the risk dominant equilibrium; see Kandori et al. (1993) and Young (1993).

In the style of lucid dialogue, Selten [1991a] reviews the above three justifications of Nash equilibrium and, more generally, he discusses the relevance of these approaches for actual human economic behavior. He concludes that we know disappointingly little about actual behavior, that we should do empirical research and that we need theories of bounded rationality. Such theories should not deal just with the cognitive bounds on human rationality but also with the motivational bounds, i.e., with the problem that even if a game has a clear and simple solution, well-informed players may not be motivated to play it. This latter paradox is well known from the finitely repeated prisoners' dilemma. This game has a clear

backwards induction solution (defect in each period), but even if all participants are well informed game theorists, an outside observer will see cooperation emerge provided the horizon is long enough. Selten [1980b] discusses a version of this paradox — the chain store paradox — and the second half of that paper proposes an informal three-level theory of bounded rationality to resolve it. Quite interestingly, the referee of a leading journal asked Selten to eliminate this second half of the paper and when he refused the paper was rejected. Only later did the paper appear elsewhere.

It is also interesting to note that Nash had already commented on the prisoners' dilemma paradox. When confronted with the outcome of a 100 times repeated prisoners' dilemma experiment conducted by Merrill Flood, Nash responded that the "grim punishment strategies" (cooperate as long as the other does so, otherwise defect forever thereafter) are "very nearly at equilibrium and in a game with an indeterminate stop point or an infinite game with interest on utility it is an equilibrium point". Furthermore, Flood (1959) quotes Nash as follows:

Since 100 trials are so long that the Hangman's Paradox cannot possibly be well reasoned through on it, it's fairly clear that one should expect an approximation to this behavior which is most appropriate for indeterminate end games with a little flurry of aggressiveness at the end and perhaps a few sallies, to test the opponents mettle during the game. (fn. 11)

The behavior that Nash describes is indeed the type that has been observed many times in the laboratory. The observed behavior can be explained either on the basis of a full rationality model incorporating incomplete information *à la* Harsanyi, as in Kreps et al. (1982), or by means of a behavioral learning model, as in Selten and Stoecker [1986a]. Experimental economics, an important offspring of game theory, will help us to clarify which explanation is most appropriate and to better understand differences in the strategic behavior of *homo sapiens* and *homo oeconomicus*.

VIII. Conclusion

von Neumann and Morgenstern (1944) proposed a new approach to economic analysis, game theory, for which they expressed certain desiderata and concerns cited here. As the quotations show, their goal was to find a mathematical general theory for strategic interaction between rational decision makers, a theory that is not self-refuting, and which allows for the possibility of elements of irrationality. The capacity to handle games of incomplete information was also a desideratum, which, however, they left for future research. von Neumann and Morgenstern

succeeded in obtaining a non-self-refuting theory for rational decision making in a special class of interactions, zero-sum two-player games.

John Nash generalized this theory to an arbitrary (finite) number of players with arbitrary von Neumann-Morgenstern preferences. Nash showed that his more general theory has a non-empty solution set in all finite games, a set that coincides with von Neumann's and Morgenstern's solution set when applied to the subclass of zero-sum two-player games. Hence, Nash succeeded in meeting the desideratum of generality in these two respects. However, the theory could still not handle games of incomplete information, nor did it allow for the possibility of irrationality. John Harsanyi envisaged a general method whereby incomplete information could be handled. He showed how a game of incomplete information can be transformed into an equivalent game of complete but imperfect information, a game to which the theory can be applied. In this way, the domain of the theory was now extended to include games of incomplete information. Reinhard Selten broadened the theory by allowing the possibility of irrationality. His perfection criterion selects those Nash equilibria which are robust to some small element of irrationality. As a by-product, he obtained subgame perfection, a criterion which is widely used to discard Nash equilibria that involve non-credible threats. von Neumann and Morgenstern's desideratum of having a complete, fully determined, theory was established by Nash only for a special case and by heuristic methods. In joint work, that was inspired by Nash, Harsanyi and Selten succeeded in finding a coherent, fully determined theory of rational interaction in games.

Game theoretic reasoning is by now established as the foundation of major parts of modern economics. Game theory is also rapidly intruding on other disciplines such as political science and sociology (in addition to biology). In view of the critical importance of having a coherent and general theory of strategic interaction, and the conceptual and technical difficulties associated with the task of establishing such a theory, the works of Nash, Harsanyi and Selten constitute milestones in the development of a common methodology within all the social sciences.

References

- Arrow, K. J. & Debreu, G.: Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90, 1954.
- Aumann, R. J.: Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, 67–96, 1974.
- Aumann, R. J.: What is game theory trying to accomplish? In K. Arrow & S. Honkapohja (eds.), *Frontiers of Economics*, Basil Blackwell, Oxford, 28–100, 1987a.
- Aumann, R. J.: Game theory. In J. Eatwell, M. Milgate & P. Newman (eds.), *The New Palgrave Dictionary of Economics*, Blackwell, Oxford, 460–82, 1987b.

- Bellman, R.: *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
- Chatterjee, K. & Samuelson, W.: Bargaining under incomplete information. *Operations Research* 31, 835–51, 1983.
- Cournot, A.: *Researches into the Mathematical Principles of the Theory of Wealth*. MacMillan, New York (first published in 1838), 1987.
- Debreu, G.: Economies with a finite set of equilibria. *Econometrica* 38, 387–92, 1970.
- Flood, M.: Some experimental games. *Management Science* 5, 5–26, 1959.
- Gibbons, R.: *A Primer in Game Theory*. Harvester Wheatsheaf, New York, 1992.
- Hayek, F. A.: Economics and knowledge. *Economica* IV, 33–54, 1937.
- Kandori, M., Mailath, G. J. & Rob, R.: Learning, mutation and long run equilibria in games. *Econometrica* 61, 29–56, 1993.
- Kohlberg, E. & Mertens, J.-F.: On the strategic stability of equilibria. *Econometrica* 54, 1003–37, 1986.
- Kreps, D. M.: *Game Theory and Economic Modeling*. Oxford University Press, Oxford, 1990a.
- Kreps, D. M.: *A Course in Microeconomic Theory*. Harvester Wheatsheaf, New York, 1990b.
- Kreps, D. & Wilson, R.: Sequential equilibria. *Econometrica* 50, 863–894, 1982.
- Kreps, D., Milgrom, P., Roberts, J. & Wilson, R.: Rational cooperation in the finitely repeated prisoner's dilemma. *Journal of Economic Theory* 27, 245–52, 1982.
- Kuhn, H. W.: A simplified two-person poker. In H. W. Kuhn & A. W. Tucker (eds.), *Annals of Mathematics Studies* 24, Princeton University Press, Princeton, NJ, 1950.
- Kuhn, H. W.: Extensive games and the problem of information. *Annals of Mathematics Studies* 28, 193–216, 1953.
- Leonard, R. J.: Reading Cournot, reading Nash: The creation and stabilisation of the Nash equilibrium. *Economic Journal* 104, 492–511, 1994.
- Luce, R. D. & Raiffa, H.: *Games and Decision*. Wiley, New York, 1957.
- Mathematica: Reports of the U.S. Arms Control and Disarmament Agency, prepared by Mathematica Inc., Princeton, NJ, ST-80, 1966; ST-116, 1967; ST-140, 1968.
- Maynard Smith, J. & Price, G.: The logic of animal conflict. *Nature* 246, 15–18, 1973.
- Mertens, J.-F.: Stable equilibria — A reformulation, Part I, Definition and basic properties. *Mathematics of Operations Research* 14, 575–625, 1989.
- Mertens, J.-F. & Zamir, S.: Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory* 14, 1–29, 1985.
- Morgenstern O., Vollkommene Voraussicht und Wirtschaftliches Gleichgewicht. *Zeitschrift für Nationalökonomie* VI (3), 337–57, 1935. English translation in A. Schotter (ed.), *Selected Economic Writings of Oskar Morgenstern*, New York University Press, New York, 1976.
- Myerson, R. B. & Satterthwaite, M. A.: Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29, 265–85, 1983.
- Neumann, J. von: Zur Theorie der Gesellschaftsspiele. *Mathematical Annals* 100, 295–320, 1928.
- Neumann, J. von & Morgenstern, O.: *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ (first edition 1944), 1947.
- Rubinstein, A.: Perfect equilibrium in a bargaining model. *Econometrica* 50, 97–109, 1982.
- Schanuel, S. H., Simon, L. K. & Zame, W. R.: The algebraic geometry of games and the tracing procedure. In R. Selten (ed.), *Game Equilibrium Models, Vol. 2: Methods, Morals and Markets*, Springer Verlag, 9–43, 1991.
- Schelling, T. C.: *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.
- Shapley, L. S.: Some topics in two-person games. In M. Dresher, L. S. Shapley & A. W. Tucker (eds.), *Advances in Game Theory*, Princeton University Press, Princeton NY, 1964.

- Ståhl, I.: *Bargaining Theory*. Economic Research Institute, Stockholm School of Economics, 1972.
- Taylor, P. & Jonker, L.: Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–56, 1978.
- Vickrey, W.: Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16, 8–37, 1961.
- Young, P.: The evolution of conventions. *Econometrica* 61, 57–84, 1993.
- Zermelo, E.: Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. In E. W. Hobson & A. E. H. Love (eds.), *Proceedings of the Fifth International Congress of Mathematicians*, Vol. 2, Cambridge University Press, Cambridge 1912.