3

Ariel Rubinstein

A subjective perspective on the interpretation of economic theory

In keeping with the general framework of this book, I will confine myself to a short summary of my own subjective perspective on economic theory. It is a perspective that began to take shape in my undergraduate days at the Hebrew University of Jerusalem, where I started my academic life as a student of mathematics. The university had a programme called 'Mathematics and combinations' which allowed us, besides concentrating on mathematics, to take courses in economics or any other academic field of our choosing. The combination of mathematics and economics was quite rare. As a group of students enrolled in this programme, we found ourselves fascinated by the sharp intellect and brightness of our mathematics teachers. The economics courses were easy and boring in comparison. We felt economics was intellectually inferior to the 'queen of sciences'. We were not, however, mathematics fanatics. In the cafeteria, where we spent many happy afternoons engaged in discussion, we found ourselves departing from the mathematical symbols in the search for metaphors for the highly abstract topological and measure-theoretical concepts. We would attempt to give these notions a real-life verbal interpretation. At the time we did not know of any mathematical model in the social sciences, and economics was still a combination of boring verbal material and unbearable calculations of derivatives. We felt intuitively, however, that something lay beyond the symbols.

On graduating to more advanced courses we were fortunate to meet two great economists, Bob Aumann and Menachem Yaari. Both these scholars proved to be thorough thinkers and firmly believed in the value of interpreting models. They provided us with standards for using mathematical tools in

The editor, A. Heertje, provided the encouragement for this essay. I am grateful to David Cooper for his valuable comments.

economics, and used mathematics rigorously. They insisted that definitions be well defined and proofs properly proved. By changing words for mathematical symbols we basically distance ourselves from the layman. Such a move can be justified if we insist that mathematics be used to ensure accuracy. Aumann and Yaari taught us that in the absence of a proper set-up for the model, these objectives cannot be achieved. Many economists still reject this view and regard such requirements as a mathematical conspiracy. The use of mathematics, however, is not necessarily beneficial. As a former student of mathematics I myself was guilty, being infatuated with mathematical thoroughness as a criterion for good economic theory. Occasionally a theorem is responsible for a surprising discovery but often it helps cover up a deficient 'economic content'.

The power of elementary mathematics first struck me in a second-year undergraduate course in mathematical logic. I enjoyed the course more than any other I have ever taken and owe much of my perspective on economic theory to it. The teacher, Saaron Shelach, made the course so difficult and intensive that it was all I could do to copy notes from the blackboard during class. However, the long days and nights I spent trying to decipher the symbols and poring over the challenging homework remain among the happiest moments of my academic life. I was fascinated by the simplicity and ingenuity of the definitions and the strength of the conclusions. There was a sense of 'touching the heart' of human reasoning.

Today, 12 years after finishing my PhD, I am still puzzled by the question of how mathematical theorems are relevant to real-life reasoning. This topic is probably the most interesting in economic theory. Notwithstanding certain doubts, I am still fascinated by the same question that occupied and enlightened me during my happy undergraduate days. This question concerns the almost magical connection between the formal definitions and theorems on the one hand and statements of natural language on the other hand. I find both the search for the interpretations and that for the correct language of economic theory very exciting. Indeed, these issues constitute the main drive of my personal research.

The key word in human behaviour is in my opinion the word 'language' and this chapter is built around this word. The chapter is a collection of independent comments organized as follows:

1. *Language and decision theory.* In the first part (based on Rubinstein, 1978) it is argued that decision-making language should be a part of decision theory.
2. *Language for stating theorems about theorems.* Mathematical logic deals with language in which theorems are stated, and this part (based on Rubinstein, 1984) is a demonstration of the possibilities of using mathematical logic methods to prove theorems about theorems.
3. *The language of bargaining theory: preferences versus utilities.* In this part (based on Rubinstein, Safra and Thomson, 1990) bargaining theory is used to demonstrate the importance of the right choice of language when stating a theory.

4. *The choice of names for concepts: what is a 'strategy'?* In this part (based on Rubinstein, 1991) it is argued through discussion of the concept of 'strategy' in an extensive game, that the choice of name applied to a concept can lead to serious difficulties.
5. *The interpretation of equilibrium strategies in sequential games.* Finally, it is argued that the interest in game-theoretic results lies in the possibility of interpreting the strategies and that 'automata' are a useful tool in the rigorous interpretation of a strategy in sequential games. This part uses ideas from Rubinstein (1986) and Rubinstein and Wolinsky (1990).

1. Language and decision theory

Decision theory is the most primitive theory in which the word 'language' can be expected to appear. In most economic models, a decision-maker is described simply by a preference relation (or equivalently a utility function). We permit preferences to be quite unrestricted, and in particular we allow them to be non-verbal. Rubinstein (1978, p. 2) tries 'to exemplify the importance of including the element of language in discussions within the theories of social choice, utility and measurement'. My argument was that: 'We instinctively justify to ourselves any decision and try to rationalize it. We are inclined to formulate judgments in words and to justify them in words' (p. 16). The need to define a preference in words is natural in cases where the decision-maker is a group of agents and decision-making involves communication among members of the group. Even when the decision-maker is an individual, he tends to justify his behaviour by giving reasons so that the preferences must be expressible in some kind of language. We must therefore look at the implications of the restrictions imposed by language on the set of relevant preferences. Whether something is 'definable' depends on the decision-maker's language and therefore the study of the connection between language and the admissible preferences should be at the heart of decision theory.

The following is one of the three examples which were presented in Rubinstein (1978). A youth arrives in a strange town. He is about to receive two offers of friendship. The information he receives about each of the girls consists of a list of boys she has dated over the last n days. Given that he is a stranger in town, let us assume that all he can glean from any two given names on the list is whether they are identical. This language is called the 'pure language with equality'. The atomic formulae in this language are formulae of the type $z_1 = z_2$, where z_1 and z_2 are symbols from among $\{x_1, \dots, x_n, y_1, \dots, y_n\}$. A formula is either an atomic formula or a string of symbols which is constructed inductively by the rules: if ϕ and ψ are formulae then '$-\phi$', 'ϕ and ψ', 'ϕ or ψ', '$\forall\phi(x)$', and '$\exists\phi(x)$' are all formulae as well.

What preferences can be stated in this simple language? That is, which formulae can induce a preference suitable for all possible 'worlds' in which our stranger is likely to find himself? Or formally, what are the definable preferences where definability of the preference is taken as the existence of a formula $\phi(x_1, \ldots, x_n, y_1, \ldots, y_n)$ so that the stranger prefers the girl with the list $(a_1, \ldots, a_n)$ over the girl with the list $(b_1, \ldots, b_n)$ iff the formula ϕ is true where each variable x_i is replaced by a_i and each variable y_i is replaced by b_i. To answer this question, define $E(a_1, \ldots, a_n)$ to be the partition of $\{1, \ldots, n\}$ in which i and j are in the same partition iff $a_i = a_j$. Thus, if the girl is unstable in her relationships and all a_i are distinct $E(a_1, \ldots, a_n)$ is the finest partition $\{\{a_1\}, \ldots, \{a_n\}\}$ and if she is very stable and all a_i are identical then $E(a_1, \ldots, a_n)$ is the coarsest partition $\{\{a_1, \ldots, a_n\}\}$. It is not difficult to verify that the only definable preferences are those which are induced by orderings of partitions of $\{1, \ldots, n\}$, so that

$$(a_1, \ldots, a_n) > (b_1, \ldots, b_n)$$

if $E(a_1, \ldots, a_n)$ is preferred to $E(b_1, \ldots, b_n)$. In other words, any information about the equality of some a_i to b_j is ignored and only the patterns of stability matter.

The above was included in a paper which was one of my very first. The paper has never been accepted for publication, probably because the specific results were not striking and the linguistic constraints which I studied were remote from the constraints imposed by the natural language. Nevertheless, the general idea that language plays an important role in determining the decision-makers' patterns of behaviour is of interest. This idea is completely missing in current economic theory.

Notice that the decision-maker uses the 'pure language with equality' which has the remarkable property (see Robinson, 1963) that definability is equivalent to 'definability without quantifiers', i.e. every formula has an equivalent formula without quantifiers. This means that the definability in that example is equivalent to the requirement that given four vectors $(a_1, \ldots, a_n)$, $(b_1, \ldots, b_n)$, $(c_1, \ldots, c_n)$ and $(d_1, \ldots, d_n)$, we require that if for all i and j, $a_i = a_j$ iff $c_i = c_j$, $b_i = b_j$ iff $d_i = d_j$ and $a_i = b_j$ iff $c_i = d_j$ then the vector $(a_1, \ldots, a_n)$ is preferred to $(b_1, \ldots, b_n)$ iff $(c_1, \ldots, c_n)$ is preferred to $(d_1, \ldots, d_n)$. This requirement is close in spirit to the neutrality assumption so common in social choice literature.

This connection is useful for reinterpreting some of the results of social choice theory. In Arrow's framework the social ordering has to be based on the preferences of the n individuals who make up the society. Let P be the symbol for the social ordering and P_i be the symbol for individual i's preference. The social choice (strong) neutrality requirement made in social choice theory (and implied by Arrow's independence of irrelevant alternatives) states that for any two pairs of social alternatives a, b and c, d, if for all i, aP_ib iff cP_id, then aPb iff cPd. The linguistic way of expressing the axiom is that the social relation has to be 'definable without quantifiers', i.e. P should be defined by a formula without quantifiers, $\phi(x_1, x_2)$, in the language in which the atomic formulae are of the type $z_1P_iz_2$ where the z_i are variable names. (For a formal proof of this fact see Rubinstein, 1984.) Thus, an alternative formulation of Arrow's impossibility theorem is that given a world with at least three individuals and three social alternatives, any social welfare function which satisfies definability without quantifiers and the Pareto-optimality requirements is dictatorial.

I find the linguistic interpretation of the neutrality type of axiom appealing. An even more attractive axiom is definability; a social ordering must be defined by a formula (not necessarily without quantifiers). Definability cannot replace definability without quantifiers in Arrow's theorem. Consider, for example, the social welfare function which chooses the majority rule ordering in the case where it induces an ordering or coincides with individual 1's preference otherwise. This function is not dictatorial and is defined in a language which includes the names of the individualistic preferences only. However, in the definition we must use quantifiers (in order to state the sentence 'if the majority rule induces an ordering'). To my knowledge the characterization of the definable preferences is still an open question.

2. Language for stating theorems about theorems

Economic theory may be viewed as the study of models which are used by economists: as such, it is not about economics but about models. In order to demonstrate this point let us continue the discussion of social choice theory begun in Rubinstein (1984). I view that paper as one of my better efforts and thus was frustrated by the rejection letters I received.

Social choice can be divided into multi-profile and single-profile theorems. A multi-profile theorem is about functions which assign a social ordering to every profile of preferences within a large set (all *possible societies*, not only the existing one). Such a theorem relies on a 'glue axiom', i.e. an axiom which requires dependency (or consistency) between the way that the social ordering treats a pair of alternatives in two different profiles where the two profiles treat the alternatives similarly. The most famous axiom of this sort is the independence of irrelevant alternatives. On the other hand, a single profile theorem refers to a single profile (the existing society). The aim is to attach an ordering to the society which will be sensitive to individuals' preferences. A typical axiom here is the neutrality axiom which requires that if each individual's preference over the alternatives *a* and *b* is 'the same' as his preferences between *c* and *d*, then the social ordering treats *a* and *b* like it treats *c* and *d*.

Social choice theory includes many 'single-profile analogues' to theorems which were proved first in the multi-profile framework. This motivated the following statement from Sen (1977, p. 1564): 'As a result of these important contributions it is now clear that the standard inter-profile collective choice results have exact intra-profile counterparts...'. Sen's conclusion is an assertion about propositions. Its formal statement requires a definition of the term 'analogue theorem'. That is where mathematical logic comes in.

Let M be the class of all possible profiles. A social function (SF) is one which assigns a binary relation to any profile in M. Let β and δ be two formulae without quantifiers in a language which includes only the symbols $P, P_1, \ldots, P_k$. Let T^* be the following theorem: if an SF satisfies strong neutrality and the proposition $\alpha = \forall v_1, \ldots, v_k \beta(v_1, \ldots, v_k)$, then it satisfies the proposition $\zeta = \forall v_1, \ldots, v_k \delta\ (v_1, \ldots, v_k)$. To identify that Arrow's impossibility theorem can be written in the scheme of T^* with $k=3$, take α to be the sentence

$$\forall v_1 v_2 v_3 \left[\left(\bigcap_{i=1,\ldots,n} v_2 P_i v_1 \rightarrow v_2 P v_1\right)\right]$$

(which expresses Pareto-optimality) in conjunction with the requirement that P is an ordering and have

$$\beta = \bigcup_{i=1,\ldots,n} \forall v_1 v_2 (v_1 P_i v_2 \rightarrow v_1 P v_2)$$

which expresses the statement that there is a dictator.

Apparently the number k plays a critical role in the formulation of a theorem about the single-profile analogues. The single-profile analogue holds only if the single profile satisfies a 'richness property', R_k, which states: every formula without quantifiers with k variables which is satisfied by some k alternatives in some profile in M is satisfied by *some* assignment of k elements in the single profile.

Now we can state a theorem about theorems: if the theorem T^* is valid for M and if the single profile satisfies R_k (where k is the same k which appears in the statement of T^*!), then the theorem T^* is true in the single profile as well.

Rubinstein (1984) includes an example of a proposition in which α does not have the structure as it exists in T^* and for which the single-profile analogue does not exist. Thus, we conclude that the logical structure of the multi-profile proposition is a key for the existence of a single-profile analogue.

The above points to a potentially interesting line of research: searching for a formal expression of the often discussed intuitions about the connections between different models on the basis of the logical structures of the theorems in which we are interested. At the moment I am not familiar with any work in this direction.

3. The language of bargaining theory: preferences versus utilities

The language of a model affects its possible interpretations. Nash bargaining theory provides an interesting example of the effect of an unsuitable choice of primitives. The discussion in this section follows Rubinstein, Safra and Thomson (1990).

The primitives of Nash's (two-person) bargaining theory are the 'feasible set', S, and a 'disagreement point', d. Each element of S gives the utility levels reached by the two agents at one (or more) of the possible agreements. The utilities are understood to be *von Neumann–Morgenstern utilities* in that they are derived from preferences over lotteries which satisfy the expected utility assumptions. A bargaining *solution* is a function which assigns a unique pair of utility levels to each problem $\langle S, d\rangle$ taken from some domain. Nash showed that there is a unique solution satisfying the following four axioms: *invariance to positive affine transformations* (INV), *symmetry* (SYM), *Pareto-optimality* (PAR) and *independence of irrelevant alternatives* (IIA). The unique solution is the Nash solution, i.e. the function N defined by

$$N(S, d) = \arg\max\{u_1 - d_1)(u_2 - d_2) \mid (u_1, u_2) \in S \text{ and } u_i \geqslant d_i \text{ for both } i\}$$

The very simplicity of this formula is in itself an attractive feature and is responsible for the widespread application of the solution. However, my problem with the above formula is that I simply do not understand its meaning. What is a product of two von Neumann–Morgenstern utility numbers and what is the meaning of the maximization of that product?

A good bargaining solution should have an attractive verbal definition. The search for a more meaningful definition of the Nash solution leads to a switch from utility language to alternatives-preferences language. A Nash problem, $\langle S, d\rangle$, is replaced by $\langle X, D, \geqslant_1, \geqslant_2\rangle$ where X is a set of feasible (deterministic) alternatives described in physical terms, D is the disagreement alternative and $\geqslant_1$ and $\geqslant_2$ are preferences defined on the space of lotteries in which the prizes are D and the elements of X.

Recall that we are looking for an alternative definition of the Nash solution, one which only uses the terms 'alternative', 'disagreement' and 'preference' and which avoids the term 'utility'. We look for an explicit definition which specifies the outcome of a particular problem directly in terms of the problem without referring to consistency with the outcomes suggested by the solution of other problems.

For the alternative definition, denote by px the lottery which gives x with probability p and D with probability $1 - p$.

DEFINITION An (ordinal)-Nash solution outcome for the problem $\langle X, D, \geqslant_1, \geqslant_2\rangle$ is an alternative y^* such that for all $p \in [0,1]$ and for all $x \in X$ and i, if $px >_i y^*$ then $py^* \geqslant_j x$.

Thus we interpret the solution as a convention which assigns to every bargaining problem an outcome with the following property: assume that the players perceive that whenever they raise an objection to an alternative, they face a risk that the negotiations will end in disagreement. If it is worth while for one of the players to make a demand for an improvement upon the convention, which may cause a breakdown of the negotiations, then it is optimal for the other player to reject the demand and to insist on following the convention even when taking into account the possibility of negotiations breaking down.

It is interesting to note that the above definition is close in spirit to an idea suggested by Zeuthen (1930) who was the first to build a theory in which negotiators bear in mind the risk of a breakdown in negotiations.

The first thing to notice about the ordinal definition is that for the expected utility case it coincides with the definition of the (utility)-Nash solution. Basically, it follows from the fact that the alternative y^* satisfies

$u_1(y^*)u_2(y^*) \geqslant u_1(x)u_2(x)$ for all x for both i

if and only if

for both i, for all $x >_i y^*$ for all $p \leqslant 1$, if $p > u_i(y^*)/u_i(x)$ then $p \geqslant u_j(x)/u_j(y^*)$

if and only if

for all i and x, $px >_i y^*$ implies $py^* \geqslant_j x$

The switch to the alternatives-preferences language allows a restatement of the entire Nash theory. In particular, it is possible to translate the axioms into more natural language and to derive the Nash characterization theorem. As long as we fix the set of alternatives (and only allow a variation of the preferences) the Invariance of Affine Transformations (IAT) axiom is equivalent to the requirement that rescaling of the utilities should not affect the alternative predicted by the solution. Once we switch to the alternatives-preferences language, IAT becomes redundant. Nash's axioms PAR and SYM can easily be translated. The main difficulty is to restate IIA: if a^* is the solution outcome of the problem $\langle T, d\rangle$ and is a member of a set S which is a subset of T, then a^* is also the solution outcome of $\langle S, d\rangle$. As has often been emphasized, this justification of IIA fits in with a normative theory, where the solution concept is intended to reflect the social desirability of an alternative. When bargaining is viewed as a strategic interaction of self-interested bargainers the IIA axiom is questionable. The following is an alternative statement of IIA which does not require a comparison between problems with different sets of alternatives (we use the letter F to denote a solution):

IIA: Let $F(\geqslant_1, \geqslant_2) = y^*$ and let $\geqslant_i'$ be a preference which agrees with $\geqslant_i$ on the set of deterministic agreements, X, such that:

1. For all x such that $x \geqslant_i y^*$, if $px \sim_i y^*$ then $px \leqslant_i' y^*$.
2. For all x such that $x \leqslant_i y^*$, if $x \sim_i qy^*$ then $x \sim_i' qy^*$.

Then $F(\geqslant_i, \geqslant_j) = F(\geqslant_i', \geqslant_j)$.

The switch of agent i's preference from $\geqslant_i$ to $\geqslant_i'$ reflects his increased apprehension towards the risk of demanding alternatives which are *better* than the outcome y^*. Though player i still prefers x to y^* he is less willing to risk demanding x. The axiom captures an intuition that the bargaining solution outcome y^* should be defendable against possible objections. The change in player i's preference, which is described in the axiom, makes player i 'less eager' to object and does not change the intensity of player j's objections. Thus, the change in the preference 'should not' change the bargaining outcome.

The ordinal definition allows us to extend Nash theory to preferences beyond those which satisfy the expected utility theory. In recent years there has been a growing interest in non-expected utility theories of decision-making under uncertainty since they explain a wide range of behaviour patterns and experimental results that are inconsistent with expected utility theory (see Machina, 1987). Extending Nash (1950) we could show that for a certain class of preferences (which includes all expected utility preferences) the Nash solution is the unique solution which satisfies the axioms PAR, SYM and IIA. A by-product of the definition is a better understanding of classical results such as the proposition that the more risk-averse the player, the worse his outcome in the bargaining.

Another by-product of the definition is a better understanding of the connection between Nash's theory and the strategic alternating offers model (see Rubinstein, 1982). The existence of a connection between these two models was first cited by Binmore (1987). When Binmore presented his result it was like a puzzle. Whereas the Nash bargaining theory reflects the attitude of the players towards risk, Rubinstein (1982) dealt with time preferences. However, in Binmore, Rubinstein and Wolinsky (1986) we equalized the primitives by looking at a version of the infinite alternating offers model where the players do not have time preferences but at the end of each period there is a probability $1 - p > 0$ of breakdown. For the expected utility case the model has a unique subgame perfect equilibrium characterized by two alternatives $x^*(p)$ and $y^*(p)$ satisfying $px^*(p) \sim_1 y^*(p)$ and $py^*(p) \sim_2 x^*(p)$. Player 1 always offers $x^*(p)$ and accepts any alternative y such that $y \geqslant_1 y^*(p)$, while player 2 always offers $y^*(p)$ and accepts any alternative x such that $x \geqslant_2 x^*(p)$.

Given the ordinal Nash bargaining solution, the connection between the alternating offers equilibrium outcomes, $x^*(p)$ and $y^*(p)$, and $N(\geqslant_1, \geqslant_2)$ is clear: where $p \to 1$, both $x^*(p)$ and $y^*(p)$ converge to $N(\geqslant_1, \geqslant_2)$.

To summarize, the utility language allows the use of geometrical presentations and facilitates analysis; however, the parametric presentation results in an unnatural statement of the solution and axioms and the judgement and interpretation of the axioms and bargaining solutions are made more difficult. The difficulties are even more severe when 'technical' assumptions (such as continuity and differentiability) are made. The switch to the alternatives-preferences language allows a more natural statement of the Nash solution. It enables us to extend the definition to non-expected utility preferences and helps us to understand better certain well-known results. Making this switch in other areas of economics and game theory may also prove beneficial.

4. The choice of names for concepts: what is a 'strategy'?

Choosing a name for a technical term satisfies our desire to link economic theory with the real world. However, it also creates the risk that a particular interpretation which is correct in one context will be incorrectly applied in another. An example of such a case is the use of the word 'strategy' in game theory. Rubinstein (1991) discusses the issue more generally; here I make do with a summary of the argument for extensive games.

A strategy in game theory is usually interpreted as 'a plan of action', 'a complete description of how a player intends to play a game, from beginning to end' or 'a set of instructions'. This interpretation is consistent with the use of the term 'pure strategy' in a normal form game. However, is it appropriate in the context of extensive games? In an extensive game, a player's strategy is required to specify an action for each node in the game tree at which the player has to move. Accordingly, a player has to specify an action for every sequence of events which is consistent with the rules of the game. In games which require a player to make at least two consecutive moves (and most of the games which have been analysed recently in economic theory fall into this category), a strategy must have its actions specified even after histories which are inconsistent with the player's own strategy. For illustration, consider the following two-player game form:

```
  C      C      C
1 ------ 2 ------ 1 -----
|S      |S      |S
```

According to the natural definition of 'strategy' as a complete 'plan of action', player 1 is required to specify his behaviour, 'Continue' or 'Stop', at the initial node and, if he plans to 'Continue', to make provisional plans for his second decision node in the event that player 2 chooses C. However, the game-theoretic definition of strategy requires player 1 to specify his action at the second decision node, even if he plans to 'Stop' the game at the first node.

Why does the notion of strategy as used by game theorists differ from a 'plan of action'? If we were only investigating Nash equilibria of extensive games, then the game-theoretic definition would indeed be unnecessarily broad. The broad definition is, however, necessary for testing the rationality of a player's plan, both at the beginning of the game and at the point where he must consider the possibility of response to an opponent's potential deviation (the subgame perfect idea). Returning to the example above, assume that each player plans to choose 'Stop' at his first decision node. Testing the optimality of player 2's plan following player 1's deviation, requires player 2 to specify his expectations regarding player 1's plan at his second decision node. The specification of player 1's action after both players have chosen C provides these expectations and has to be interpreted as what would be player 2's (as opposed to player 1's) *belief* regarding player 1's planned future play, should player 1 decide to deviate from what was believed to be his original plan of action. Thus, a strategy encompasses not only the player's plan but also his opponent's expectations in the event that he does not follow that plan. Thus, an equilibrium strategy describes a player's plan of action, as well as those considerations which support the optimality of his plan (i.e. preconceived ideas concerning the other player's plans) rather than being merely a description of a 'plan of action'.

Interpreting a player's strategy after a deviation as the expectations of the other players about his future behaviour, makes it problematic to speak of a 'choice of strategy'. Player 1 does not choose player 2's belief. This observation has a serious impact on many of the game-theoretic assumptions. Consider, for example, the sequential bargaining literature in which the authors *assume* that strategies are stationary in the sense that a player's offer and response (to offers made by the other player) must be independent of the history of the game. This literature presents stationarity as an assumption of simplicity of behaviour. Consider, for example, player 1's strategy: 'Demand 50 per cent of the surplus and reject any offer which gives you less than 50 per cent, independent of what has happened in the past.' This strategy is simple in the sense that player 1 plans to make the same offer and make the same responses independently of how player 2 has reacted in the past. However, this strategy also implies that player 2 believes that player 1 would demand 50 per cent of the surplus even if player 1 demanded 60 per cent of the surplus in the first, let us say, 17 periods of bargaining. Thus, stationarity, as stated in sequential bargaining theory, means not only simplicity but also passivity of beliefs. This is strange, especially if we assume simplicity of behaviour. If player 2 believes that player 1 is constrained to choose a stationary plan of action, then player 2 should believe (after 17 repetitions of the demand of 60 per cent) that player 1 will continue to demand 60 percent. Thus, assuming passivity of beliefs eliminates a great deal of what sequential games are intended to model, namely, the changing pattern in players' behaviour and beliefs, as they accumulate experience.

5. The interpretation of equilibrium strategies in sequential games

Many of the major contributions to economics and game theory are concerned with types of games which have similar structures to repeated games. Much of the research in this area aims to prove what is called 'folk theorems'. This (inappropriate) name is given to theorems which state that under certain conditions, nearly all reasonable pay-off vectors can be sustained as equilibrium pay-off vectors. Thus, according to these theorems, pay-off vectors which are socially desirable can be maintained in equilibrium if the players have in mind long-term considerations; however, the model lacks predictive power.

In my opinion, the main achievement of these models is in clarifying the logic behind social institutions associated with interactions over the long term.

Characterizing the exact set of equilibrium pay-off vectors is less exciting than discussing the equilibrium *strategies*, since the verbal content of the equilibrium lies in the strategies and not in the pay-off vectors. Most literature on repeated games deals with characterizing the set of equilibrium *pay-off* vectors while the plausibility of the strategies is largely ignored. Existence theorems demonstrate unintuitive strategies. For example, the folk theorem concerning the limit of the means is sometimes proved using equilibrium in which a deviation at the nth period is met with punishment for n^2 periods. This is done for the convenience of proving the folk theorem in its maximal range. Or, in other examples, folk theorems often use equilibria which carry the element of 'punishing the punisher for not punishing', i.e. if player i 'deserves punishment' and player j does not punish player i, then player i is supposed to punish player j for not punishing him. Though I suppose there are some bizarre scenarios in which a criminal sues a policeman for not punishing him as severely as he should have, I doubt that this is a common mode of behaviour.

I myself am not blameless when it comes to proving folk theorems and ignoring the plausibility of the strategies. In my early papers on repeated games with the limit of the means (Rubinstein, 1977) and on the overtaking criterion (Rubinstein, 1979b) I did not pay enough attention to this issue. In retrospect, I believe that in order to clarify the nature of long-term interactions, we must deal with the equilibrium strategies' schemata. When discussing a scheme of strategy I am referring to its structure, stripped of the details which arise from a particular pay-off matrix. Accordingly, the value of the folk theorems is their usefulness in clarifying the rationale and credibility of codes of behaviour in which a deviator is punished for a finite number of periods before the world returns to routine behaviour.

Let me expand the discussion using two of my papers which are concerned more with the structure of strategies than with the outcome in terms of 'pay-offs': Rubinstein (1979a) is a study of the following problem: player 1 monitors player 2's behaviour. At every period player 2 can choose one of two actions, 'G' or 'B'. The B-action results (with probability 1) in a harmful accident while the G-action causes an accident only with probability $1 > p > 0$. In any particular period, player 1 has the means to punish player 2 severely enough to deter him from behaving 'badly'. However, the punishment is 'costly' for both players and player 1 looks for schemes of behaviour in which he will be able to deter player 2 effectively and cheaply. The problem can be viewed as a leader–follower situation in which player 1 leads by making a binding announcement which specifies the histories, following which player 1 punishes player 2, and player 2 responds optimally by choosing a strategy which specifies, for all sequences of events, whether he will choose 'G' or 'B'. Both players are assumed to maximize their expected 'limit of the averages' pay-offs.

The leader's dilemma is quite clear. On the one hand he wants to avoid punishing player 2 too often even if he sticks to the good behaviour mode. On the other hand, he wants to avoid player 2 taking advantage of his leniency.

The main idea of Rubinstein (1979a) (independently proposed by Radner, 1981, as well) is that player 1 can achieve his first best result by employing a strategy in which he behaves as a statistician, i.e. he keeps track of the frequency of accidents and punishes player 2 whenever the frequency exceeds $p+\alpha_t$ where $(\alpha_t)_{t=1,2,\ldots}$, is a sequence of positive numbers which is selected carefully to satisfy two conditions: (1) it converges to zero so that the follower cannot achieve a positive frequency of being 'Bad' without being punished, and (2) the convergence to zero is slow enough to allow player 1 to tolerate some of player 2's 'bad luck' accidents. The law of iterated logarithm guarantees that such a sequence exists.

In the meantime, many other strategies have been suggested in the literature to solve the leader's problem. In some, the mathematics which is required to prove their effectiveness calls for theorems which are more elementary than the law of iterated logarithm. Others rely on 'dynamic programming' techniques (see Abreu, Pearce and Staccheti, 1986, and for a survey see Pearce, 1991). The basic idea of these equilibria is the following: the equilibrium is built around two 'phases', 'A' and 'B'; in phase A the leader ignores the accidents and in phase B he does not. The switch from one phase to another is stochastic and its probabilities depend on the outcome of the one-shot situation so that the follower is motivated always to choose 'G' when he solves the 'short-term' problem in which the probabilities of continuation are determined by the transition probabilities. Personally, I find the first mechanism more plausible since it captures a type of reasoning which we observe and use in real life.

A second example is taken from Rubinstein and Wolinsky (1990). Consider a market with an indivisible good, one seller and $B > 1$ potential buyers. All buyers have an identical reservation value of 1, while the seller's reservation value is zero. The competitive equilibrium price is 1. Asher Wolinsky and my aim was to examine whether we get the competitive outcome in models with pairwise matching and bargaining. To do this, we add details about the matching and bargaining processes to the basic situation. In each period the seller is matched randomly with one buyer, and one of the two (picked with equal probability) has to make a price offer which the other either accepts or rejects. The process repeats itself until an acceptance is made. The natural competitive forces are supposed to kick in here since the buyer is under the risk of losing the potential partner while the seller is not. Indeed, the model has an equilibrium with the competitive price in which the good is sold to the first buyer which the seller meets. In that equilibrium the seller always makes the request for price 1 and rejects any lower offer. However, the model has other equilibria in which non-competitive prices prevail. Let p^* be an arbitrary price, let us say the 'fair' price according to some ethical evaluation. To defend this price we introduce the institution of the 'right' to purchase at p^*: at the beginning of the game the right is granted to one of the players i^*. The right to purchase the good for p^* is kept by the right holder until either: (1) the seller approaches buyer j with an offer above p^*, in which case the right is transferred to j (in order to neutralize the competitive forces); or (2) buyer j approaches the seller with an offer above p^* in which case the players move to the competitive regime and behave as in the equilibrium which supports price 1. In the jargon of repeated games theory, the change in the regime means that the seller punishes the buyer who deviates from the equilibrium, although he offers the seller a price higher than he would otherwise get. This is an unacceptable interpretation. We should interpret the switch as a change in the market state from a mode in which a right is given to one of the buyers to a mode which gives the competitive state. The attempt of a buyer to pay a price above p^* is taken by the agents to mean that there are two serious buyers in the market who are ready to compete and this pushes the price up to the competitive level.

In reality we do observe 'privileges' in trade. However, I am not aware of any case in which the rules of transferring a right include a provision that the right is transferred when the seller approaches another buyer. Still, I find the above equilibrium attractive as it is easily described verbally and includes components of familiar institutions.

The effort to interpret sequential games equilibrium strategies leads to the search for a formal model in which we can refer to the interpretation of the strategies more rigorously. This was part of my motivation in Rubinstein (1986) where I replaced the strategies in repeated games with a machine called a finite automaton. This structure is widely used in linguistics for analysing the structure of sentences and is a standard tool in computer sciences; sometimes it is used as a metaphor to describe the way that the human brain functions. In the context of sequential games, a player, instead of choosing a strategy, chooses a machine which implements his strategy. A machine includes four elements:

1. A set of abstract elements, each of which is a 'state of mind'.
2. An indicator of one of the states to be the 'initial state' from which the machine starts to operate.
3. An output function which assigns the action taken by the machine at each state.
4. A transition function which determines what state the machine is moving to after it receives an input, which is a piece of information about other players' actions.

Thus, an automaton is a device which organizes a player's behaviour in the game. It helps us to formulate a verbal statement of the strategy by attaching a name to each of the states. For example, let us return to the seller–buyers model discussed in the previous section. In that model the equilibrium strategies can be described as a $B + 1$ state machine. The set of states includes one state, Right(i), for each buyer i, and an extra state COMP. The state Right(i) has the interpretation of 'i has the right to purchase the good'. The state COMP is the 'competitive state'. The content of a state is determined by its output and by the rule of transition at that state. The initial state is Right(i^*). At Right(i) the seller offers p^* and accepts p^* from buyer i only; when meeting other buyers he offers a non-serious price (above the competitive price) and rejects all offers which come from anyone other than *i*. COMP is a terminal state, in the sense that once the machine gets to COMP it never leaves that state. In this state the seller demands the price 1 and accepts no less than 1. The *B* buyers' strategies are similar in their structure, having the same transition rules. (See Ben-Porath and Peleg, 1987, for presentation of results related to repeated games in the language of automata. See Osborne and Rubinstein, 1990, for presentation of results from sequential bargaining theory using this language.)

Though the automaton is used to facilitate a clear interpretation of the sequential game strategies, its main importance to economic theory is as a convenient analytical tool for introducing considerations of complexity into our models. When players choose a strategy they take into account not only their game's pay-off but also the complexity of the strategy. To do this we first need to formalize the notion of complexity of a strategy; the language of automata is a convenient vehicle for accomplishing this. First steps in this direction were taken in Rubinstein (1986) and in Abreu and Rubinstein (1988). (See also Rubinstein, 1987.) Players were required to find the trade-off between two objectives: reducing the complexity of their strategies as measured by the number of states in their machine and maximizing the repeated game pay-off. Adding the complexity consideration led to a dramatic change in the folk theorem results.

6. Concluding remarks

Let me conclude with the following remarks.

The goal of economic theory

The issue of interpreting economic theory is, in my opinion, the most serious problem facing economic theorists at the moment. Economic theory has gained a prominent place in the study of economics and achieved a remarkable influence on other social sciences as well. None the less, I find many of my colleagues somewhat apologetic about the goals and achievements of economic theory and even speak of a crisis situation. Their concern can be summarized as follows: economic theory is the part of economics which attempts to deal with the real world. It is not a branch of abstract mathematics even though it utilizes mathematical tools. Since it is about the real world we expect the theory to prove useful in achieving practical goals. But economic theory does not deliver the goods. It cannot make predictions anything like those offered by the natural sciences, and the link between economic theory and practical problems, such as how to fight inflation, is tenuous at best. Economic theory lacks a consensus as to its purpose and interpretation. Again and again, we find ourselves asking the question, 'Where is it leading?' (See Aumann, 1987; Binmore, 1983.) My belief as outlined in this chapter is that, notwithstanding its remoteness from the real world, economic theory is about the language which is used in our reasoning about social interactions.

The choice of the model's language

In economics, as in any other field of science, we look for regularities: regularity, however, depends on the language we use to describe a particular situation. Behaviour may be regular or irregular, depending on the language used. If you put blue, red and green objects in front of a subject you may find irregularity in the sense that the subject picks a different colour each time: however, his behaviour may be quite regular if we describe his problem as a choice from among left, centre and right if he always chooses left. If our vocabulary did not include the words for position, we would not be able to describe this regularity. Thus, the correct language for explaining regularities must coincide with the way in which participants perceive a situation and not the way in which analysts perceive it.

Future goals

There are two major fields of research in which language and economic theory can interact. First, tools from economic theory can be used to explain the classification systems individuals use. The function of these systems is connected to human interaction and is therefore likely to be closely linked with equilibrium analysis. Second, language is part of the reasoning process used by decision-makers. Economic models with these processes embedded in them are usually classified under the title 'bounded rationality', a field which has recently attracted wide attention. Is a new economic theory emerging which will abandon the full rationality assumption and instead focus on decision-makers' individualistic procedures? In my opinion the answer is yes, but it will remain for future research in economic theory to determine the answer.

References

Abreu, D., Pearce, D. and Staccheti, E. (1986) Optimal cartel equilibria with imperfect monitoring. *Journal of Economic Theory*, **39**, 251–69.

Abreu, D. and Rubinstein, A. (1988) The structure of Nash equilibrium in repeated games with finite automata. *Econometrica*, **56**, 1259–82.

Aumann, R. (1987) What is game theory trying to accomplish? In K. J. Arrow and S. Honkapohja (eds), *Frontiers of Economics*. Oxford: Blackwell.

Ben Porath, E. and Peleg, B. (1987) On the Folk theorem and finite automata. Research Paper No. 77, The Hebrew University of Jerusalem.

Binmore, K. (1983) Aims and scope of game theory. Mimeo, LSE.

Binmore, K. (1987b) Nash bargaining theory II. In K. Binmore and P. Dasgupta (eds), *The Economics of Bargaining*. Oxford: Blackwell, 61–76.

Binmore, K., Rubinstein, A. and Wolinsky, A. (1986) The Nash bargaining solution in economic modelling. *The Rand Journal of Economics*, **17**, 176–88.

Machina, M. (1987) Choice under uncertainty: problems solved and unsolved. *Journal of Economic Perspectives*, **1**, 121–54.

Nash, J. (1950) The bargaining problem. *Econometrica*, **28**, 155–62.

Osborne, M. and Rubinstein, A. (1990) *Bargaining and Markets*. London: Academic Press.

Pearce, D. (1991) *Repeated games: cooperation and rationality*. Cowles Foundation.

Radner, R. (1981) Monitoring cooperative agreements in a repeated principal–agent relationship. *Econometrica*, **49**, 1127–48.

Robinson, A. (1963) *Introduction to Model Theory and to the Meta Mathematics of Algebra*. Amsterdam: North-Holland.

Rubinstein, A. (1977) Equilibrium in supergames. Research Memorandum No. 25, Center for Research in Mathematical Economics and Game Theory, The Hebrew University of Jerusalem.

Rubinstein, A. (1978) Definable preferences relations–three examples. Research Memorandum No. 31, Center for Research in Mathematical Economics and Game Theory, The Hebrew University of Jerusalem.

Rubinstein, A. (1979a) An optimal policy for offenses that may have been committed by accident. In S. Brams, A. Schotter and G. Schwodiauer (eds), *Applied Game Theory*. Wurzburg: Physica-Verlag, 406–13.

Rubinstein, A. (1979b) Equilibrium in supergames with the overtaking criterion. *Journal of Economic Theory*, **21**, 1–9.

Rubinstein, A. (1982) Perfect equilibrium in a bargaining model. *Econometrica*, **50**, 97–110.

Rubinstein, A. (1984) The single profile analogues to multi profile theorems: mathematical logic's approach. *International Economic Review*, **25**, 719–30.

Rubinstein, A. (1986) Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory*, **39**, 83–96.

Rubinstein, A. (1987) The complexity of strategies and the resolution of conflict: an introduction. In B. Bryant and R. Portes (eds), *Global Macroeconomics: policy conflict and cooperation*. London: Macmillan Press, 17–32.

Rubinstein, A. (1991) Comments on the interpretation of game theory. *Econometrica*, **59**, 909–24.

Rubinstein, A., Safra, Z. and Thomson, W. (1990) On the interpretation of the Nash bargaining solution. Mimeo, Tel Aviv University.

Rubinstein, A. and Wolinsky, A. (1990) Decentralized trading, strategic behavior and the Walrasian outcome. *Review of Economic Studies*, **57**, 63–78.

Sen, A. (1977) On weights and measures: informational constraints in social welfare analysis. *Econometrica*, **45**, 1539–72.

Zeuthen, F. (1930) *Problems of Monopoly and Economic Warfare*. London: Routledge and Kegan Paul.