RATIONAL CHOICE THEORY

CRITICAL CONCEPTS IN THE SOCIAL SCIENCES

Edited by
MICHAEL ALLINGHAM

RATIONAL CHOICE THEORY

Critical Concepts in the Social Sciences

Edited by
Michael Allingham

Volume III

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2006
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Avenue, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group, an informa business

Typeset in 10/12pt Times by Graphicraft Ltd., Hong Kong
Printed and bound in Great Britain by
MPG Books Ltd., Bodmin, Cornwall

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book has been requested

ISBN10: 0-415-35751-9 (Set)
ISBN10: 0-415-35754-3 (Volume III)
ISBN13: 978-0-415-35751-7 (Set)
ISBN13: 978-0-415-35754-8 (Volume III)

Publisher's Note

References within each chapter are as they appear in the original complete work.

ACKNOWLEDGEMENTS

The publishers would like to thank the following for permission to reprint their material:

The Econometric Society for permission to reprint John C. Harsanyi, 'A general theory of rational behavior in game situations', *Econometrica*, 34, 1966, pp. 613–34.

Institute of Mathematical Studies for permission to reprint R. J. Aumann, 'Agreeing to disagree', *Annals of Statistics*, 4, 1976, pp. 1236–39.

Blackwell Publishing for permission to reprint Jane Heal, 'Common knowledge', *Philosophical Quarterly*, 28, 1978, pp. 116–31.

The American Economic Association for permission to reprint A. Schotter and G. Schwodiauer, 'Economics and the theory of games: a survey', *Journal of Economic Literature*, 18, 1980, pp. 479–527.

The Econometric Society for permission to reprint B. Douglas Bernheim, 'Rationalizable strategic behavior', *Econometrica*, 52, 1984, pp. 1007–28.

The Econometric Society for permission to reprint David G. Pearce, 'Rationalizable strategic behavior and the problem of perfection', *Econometrica*, 52, 1984, pp. 1029–1050.

Cambridge University Press for permission to reprint Ken Binmore, 'Modeling rational players: Part I', *Economics and Philosophy*, 3, 1987, pp. 179–214. Copyright © Cambridge University Press, reprinted with permission.

Cambridge University Press for permission to reprint Ken Binmore, 'Modeling rational players: Part II', *Economics and Philosophy*, 4, 1988, pp. 9–55. Copyright © Cambridge University Press, reprinted with permission.

The *Journal of Philosophy* for permission to reprint P. Pettit and R. Sugden, 'The backward induction paradox', *Journal of Philosophy*, 86, 1989, pp. 169–82.

The Econometric Society for permission to reprint Ariel Rubinstein, 'Comments on the interpretation of game theory', *Econometrica*, 59, 1991, pp. 909–24.

American Economic Association for permission to reprint A. Brandenburger, 'Knowledge and equilibrium in games', *Journal of Economic Perspectives*, 6, 1992, pp. 83–101.

American Economic Association for permission to reprint J. Geanakoplos, 'Common knowledge', *Journal of Economic Perspectives*, 6, 1992, pp. 53–82.

The American Economic Association and Philip Reny for permission to reprint Philip J. Reny, 'Rationality in extensive-form games', *Journal of Economic Perspectives*, 6, 1992, pp. 103–18.

The Econometric Society for permission to reprint Robert Aumann and Adam Brandenburger, 'Epistemic conditions for Nash equilibrium', *Econometrica*, 63, 1995, pp. 1161–80.

The American Economic Association for permission to reprint John C. Harsanyi, 'Games with incomplete information', *American Economic Review*, 85, 1995, pp. 291–303.

Disclaimer

The publishers have made every effort to contact authors/copyright holders of works reprinted in *Rational Choice Theory: Critical Concepts in the Social Sciences*. This has not been possible in every case, however, and we would welcome correspondence from those individuals/companies whom we have been unable to trace.

CONTENTS

Acknowledgements vii

26 A general theory of rational behavior in game situations 1
JOHN C. HARSANYI

27 Agreeing to disagree 26
ROBERT J. AUMANN

28 Common knowledge 30
JANE HEAL

29 Economics and the theory of games: a survey 47
ANDREW SCHOTTER AND GERHARD SCHWÖDIAUER

30 Rationalizable strategic behavior 112
B. DOUGLAS BERNHEIM

31 Rationalizable strategic behavior and the problem of perfection 137
DAVID G. PEARCE

32 Modeling rational players: Part I 162
KEN BINMORE

33 Modeling rational players: Part II 197
KEN BINMORE

34 The backward induction paradox 242
PHILIP PETTIT AND ROBERT SUGDEN

35 Comments on the interpretation of game theory 256
ARIEL RUBINSTEIN

36 Knowledge and equilibrium in games 274
ADAM BRANDENBURGER

37 **Common knowledge** 294
JOHN GEANAKOPLOS

38 **Rationality in extensive-form games** 326
PHILIP J. RENY

39 **Epistemic conditions for Nash equilibrium** 342
ROBERT AUMANN AND ADAM BRANDENBURGER

40 **Games with incomplete information** 364
JOHN C. HARSANYI

26

A GENERAL THEORY OF RATIONAL BEHAVIOR IN GAME SITUATIONS

John C. Harsanyi[1]

Source: *Econometrica* 34 (1966): 613–34.

The von Neumann-Morgenstern theory of games does not yield determinate solutions (corresponding to unique payoff vectors) for two-person variable-sum games and for *n*-person games. The present paper outlines a general theory of rational behavior in game situations which does yield determinate solutions for all classes of games. The theory is based on two classes of rationality postulates: those defining rational behavior as such, and those defining rational expectations concerning the other players' behavior. It is argued that this new approach greatly increases the possibilities for the application of game theory in economics and the other social sciences.

1

The von Neumann-Morgenstern approach to game theory yields a very satisfactory solution concept for two-person constant-sum games. But it fails in general to yield determinate solutions (i.e., solutions corresponding to a unique payoff vector) for two-person variable-sum and for *n*-person games. This greatly restricts the usefulness of their approach for economics and the other social sciences. Most real-life social situations represent *n*-person games, or possibly two-person variable-sum games where the two players' interests are not completely opposed to each other. In such social situations only a theory yielding determinate solutions can help us to *predict* or *explain* the outcome, can suggest hypotheses sufficiently specific as to allow empirical *testing*, and can furnish reasonably definite *policy recommendations* to the participants.

But we have been able to show that if one accepts certain very natural rationality postulates, then one obtains a general theory of rational behavior in game situations, yielding determinate solutions for *all classes* of finite

games (as well as for infinite games satisfying certain regularity conditions), both two-person and *n*-person, constant-sum and variable-sum, cooperative and non-cooperative, with and without transferable utility, etc.

To be sure, our own approach yields determinate predictions about empirical social behavior only if we introduce *factual assumptions* about the players' utility functions, their strategy possibilities, the information available to them, etc. In our view, game theory—like individual decision theory (utility theory), of which game theory is a generalization—should be regarded as a purely formal theory lacking empirical content. Both theories merely state what will happen if all participants have consistent preferences and follow their own preferences in a consistent manner—whatever these preferences may be. Empirical content comes in only when we make more specific assumptions about the nature of these preferences and about other factual matters (e.g., when we assume that people prefer more money to less money, or make assumptions about the strategies available to them, etc.). The advantage of our approach lies only in the fact that as soon as we have introduced the necessary factual assumptions we obtain unique predictions for all empirical social situations on the basis of the *same* general theory, without the need of theoretically unjustified *ad hoc* assumptions in each particular case. Thus, for instance, bilateral monopoly (including collective bargaining), duopoly, oligopoly, their various combinations, political power situations, etc., all become special cases of the same general theory.

More fundamentally, the purpose of our theory is to answer certain basic questions about real-life social situations, which can be answered only on the basis of a theory yielding determinate predictions. For instance, if rational individuals have a common interest in reaching an efficient solution but have opposite interests as to the specific payoff distribution to be adopted, what factors will determine whether they can actually reach an agreement yielding an *efficient* outcome? In other words, under what conditions will rational individuals be unable to reach an agreement and be driven into a wasteful *conflict situation* which would be in their mutual interest to avoid? How will the participants' *relative payoffs* (i.e., their relative bargaining power) depend on the basic independent variables characterizing a given social situation? What *coalition structure* will emerge if all participants act rationally? etc.

Let me add that the solutions our theory defines for various games are not based in any way on moral value judgments. In our view we must clearly distinguish between ethical and game-theoretical problems. *Ethics* (and also welfare economics, which for our purposes is a branch of ethics) tries to define the patterns of social behavior best serving the long-run *interests of society* as a whole. In contrast, *game theory* tries to define those patterns of social behavior that will emerge if every participant rationally pursues his own *self-interest* (or more generally pursues all values, selfish and unselfish, that happen to be included in his own utility

function) in the face of other individuals likewise pursuing their own self-interests (and/or their other personal values) in a rational manner. Of course, game-theoretical models have application also to cases where some or all players ascribe positive utility to conformity with certain moral values. But these moral preferences must always be included in each player's utility function (payoff function), rather than being used in the formal definition of the solution. Apart from greater conceptual clarity, this approach has the advantage of much greater generality because it also covers cases where the different players have dissimilar moral values or pay no attention to moral considerations at all.

Thus in our view the *general theory of rational behavior* should be subdivided into:

1. *Individual decision theory*, dealing with rational behavior by an isolated individual, and covering the cases of

   1a. Certainty,
   1b. Risk, and
   1c. Uncertainty.

2. *Ethics*, dealing with a rational pursuit of the long-run *interests of society* as a whole, and
3. *Game theory*, dealing with a rational pursuit by each individual of his own *personal interests* (as expressed by his utility function) against other individuals rationally pursuing their own personal interests—where any individual's "personal interests" may include both selfish and unselfish considerations.

In cases 1 and 2 rational behavior can be defined in terms of rather simple criteria. In case 1 (individual decision theory), recent work (e.g., Savage [**15**], Anscombe and Aumann [**1**]) has shown that if an individual's choices satisfy certain very natural rationality postulates then his behavior can be interpreted as an attempt to maximize his *expected utility* in terms of the *subjective probabilities* he assigns to alternative possibilities (Bayesian approach). In case 2 (ethics), it can be shown that if a person's moral value judgments follow certain rationality postulates then these value judgments will evaluate people's behavior in terms of a social welfare function representing the *arithmetic mean* (or equivalently the sum) of all individuals' utility functions in the society (Harsanyi [**4, 5**]). The purpose of the present paper is to indicate how rational behavior can be defined in case 3 (game theory). Our approach will be a direct generalization of Bayesian decision theory.

We feel that a clear distinction between cases 2 and 3 is very essential. A good deal of confusion in the literature could have been avoided if a clearer distinction had been made between ethical and game-theoretical

considerations, in particular between "*arbitration models*," e.g., Raiffa [**14**]; Braithwaite [**2**]; and game-theoretical "*bargaining models*" proper. The former try to define a solution satisfying certain moral "fairness" criteria. The latter try to define the solution that will emerge if all players are interested only in maximizing their own payoffs. In such "bargaining models," concessions by the players to each other will not be motivated by moral considerations but rather by the players' finding it too risky, from their own points of view, to refuse these concessions. The use of "arbitration models" where the use of "bargaining models" would have been called for has presumably resulted from the mistaken assumption that game-theoretical considerations as such cannot define a unique solution.

2

Before stating our rationality postulates we need the following definitions and notations.

By a *cooperative game* we mean a game where commitments (i.e., agreements, promises, and threats) are fully binding and enforceable. By a *non-cooperative game* we mean a game where commitments have no binding force. (The possibility or impossibility of communication does not enter into our definitions because we want to distinguish between *vocal* and *tacit* games *both* among cooperative and among non-cooperative games.)

A cooperative game can always be replaced by a non-cooperative game if we incorporate promises and threats in the strategies available to the players and use a payoff matrix making violation of such commitments result in heavy penalties (which are explicitly stated in the payoff matrix). But this procedure greatly increases the size of the payoff matrix we have to consider; and the concept of cooperative games in its usual form has some important heuristic advantages. For these reasons we shall retain the concept of cooperative games and shall not replace these games by the equivalent non-cooperative games.

Player i's pure strategies will be denoted by a_i, b_i, ... His mixed strategies (which of course include his pure strategies as special cases) will be denoted by r_i, s_i, ... The symbols r, s, . . . will be used to denote strategy n-tuples, e.g., $s = (s_1, \ldots, s_n)$. The symbols r^i, s^i, . . . will denote the strategy $(n-1)$-tuples which remain if player i's strategy is omitted from the n-tuple r, s, . . . , e.g., $s^i = (s_1, \ldots, s_{i-1}, S_{i+1}, \ldots, s_n)$. We shall write $s = (s_i, s^i)$, etc.

In a non-cooperative game the players can use only *individually* randomized mixed strategies, while in a cooperative game they can also use *jointly* randomized mixed strategies. Accordingly, when we speak of a *joint* strategy σ of all n players, in a non-cooperative game this term will always refer to some given strategy n-tuple $\sigma = s = (s_1, \ldots, s_n)$, whereas in a cooperative game it will in general refer to some probability mixture of different strategy n-tuples s, t, ...

The set of all strategies s_i available to player i will be denoted by S_i. The set S' of all strategy n-tuples available to the n-players is the Cartesian product $S' = S_1 \times \ldots \times S_n$. The set of all joint strategies σ available to the n players will be denoted by S. In non-cooperative games we can write $S = S'$. The set S^i of all strategy $(n-1)$-tuples s^i available to the $(n-1)$ players other than player i is the Cartesian product

$$S^i = S_1 \times \ldots \times S_{i-1} \times S_{i+1} \times \ldots \times S_n.$$

Let R_i be any subset of S_i. Then the mixed strategy r_i representing the equiprobability mixture of all strategies s_i in set R_i will be called the *centroid strategy* of set R_i. If R_i is a convex set then r_i will be itself an element of R_i, otherwise this need not be the case.

We denote player i's payoff function by U_i and shall write player i's payoff as $u_i = U_i(s)$. The symbol U will denote the n-tuple $U = (U_1, \ldots, U_n)$. We can regard U as a vector-valued function and can write $u = (u_1, \ldots, u_n) = U(s)$.

Let $t^i \in S^i$ be a given strategy $(n-1)$-tuple available to the $(n-1)$ players other than player i. Let $s_i^* \in S_i$ be one of player i's strategies satisfying

$$U_i(s_i^*, t^i) \geqq U_i(s_i, t^i) \quad \text{for every } s_i \in S_i. \tag{2.1}$$

Then s_i^* is called a *best reply* of player i to the other players' strategy combination t^i. If in (2.1) we can replace the $\geqq$ sign by the $>$ sign for all $s_i \neq s_i^*$ then we call s_i^* his *only best reply* to t^i. Finally let S_i^* be the set of *all* best-reply strategies s_i^* available to player i against t^i. S_i^* is always a convex set. For suppose that both $s_i^* = s_i^0 \in S_i^*$ and $s_i^* = s_i^{00} \in S_i^*$ are best-reply strategies of player i to t_i, and let $\tilde{s}_i$ be the mixed strategy $\tilde{s}_i = {}^1/_2 s_i^0 + {}^1/_2 s_i^{00}$. Then by (2.1)

$$U_i(s_i^0, t^i) = U_i(s_i^{00}, t^i) = U_i(\tilde{s}_i, t^i) \geqq U_i(s_i, t^i)$$

for every $s_i \in S_i$. Hence $\tilde{s}_i$ itself is also a best reply to t_i and $\tilde{s}_i \in S_i^*$.

Let s_i^{**} be the equiprobability mixture of *all* strategies s_i^* in set S_i^*. By the convexity of S_i^*, this strategy s_i^{**} will be itself a best reply to t^i. We shall call it player i's *centroid best reply* to t^i.

Let $s^* = (s_1^*, \ldots, s_n^*) \in S$ be a strategy n-tuple where every player's strategy s_i^* is a best reply to all other players' strategy $(n-1)$-tuple $(s^*)^i$. Then we call s^* an *equilibrium point* (cf. Nash [**12**]). If every player's strategy s_i^* is not only a best reply but is the *only* best reply to $(s^*)^i$ then we call s^* a *strong equilibrium point*. Any other equilibrium point will be called *weak*. Finally if every s_i^* is a *centroid* best reply to $(s^*)^i$ then we call s^* a *centroid equilibrium point*. Clearly, every strong equilibrium point is also a centroid equilibrium point but not conversely.

Any strategy s_i^* that is a component of some equilibrium point s^* is called an *equilibrium strategy*.

Suppose player i has no definite expectations about the strategy $(n-1)$-tuple t^i the other $(n-1)$ players will follow, but has only a *subjective probability distribution* P over all possible $(n-1)$-tuples t^i in set S^i. Let $\tilde{t}^i$ be the mixed strategy representing that particular probability mixture of all possible t^i's which corresponds to this subjective probability distribution P. Finally let $s_i^* \in S_i$ be a strategy of player i maximizing his payoff against $\tilde{t}^i$, so that

$$U_i(s_i^*, \tilde{t}^i) = \max_{s_i \in S_i} U_i(s_i, \tilde{t}^i). \tag{2.2}$$

Then s_i^* is called a *generalized best reply* to the other players' expected strategies.

Let

$$\bar{u}_i = \min_{s^i \in S^i} U_i(\bar{s}_i, s^i) = \max_{t_i \in S_i} \min_{s^i \in S^i} U_i(t_i, s^i). \tag{2.3}$$

Then $\bar{u}_i$ will be called player i's *maximin payoff*, and $\bar{s}_i$ will be called a *maximin strategy* for him.

Let $\bar{S}_i$ be the set of *all* maximin strategies $\bar{s}_i$ available to player i. $\bar{S}_i$ is always a convex set. For let $\bar{s}_i = s_i^0 \in \bar{S}_i$ and $\bar{s}_i = s_i^{00} \in \bar{S}_i$ be two maximin strategies of his, and let $\tilde{s}_i = {}^1/_2 s_i^0 + {}^1/_2 s_i^{00}$. Then by (2.3) for every $s^i \in S^i$ we have $U_i(s_i^0, s^i) \geqq \bar{u}_i$ and $U_i(s_i^{00}, s^i) \geqq \bar{u}_i$ and so also $U_i(\tilde{s}_i, s^i) \geqq \bar{u}_i$. Hence $\tilde{s}_i$ itself is also a maximin strategy and so $\tilde{s}_i \in \bar{S}_i$.

Let $\bar{\bar{s}}_i$ be the equiprobability mixture of *all* strategies $\bar{s}_i$ in set $\bar{S}_i$. By the convexity of $\bar{S}_i$, this strategy $\bar{\bar{s}}_i$ will be itself a maximin strategy. We shall call it player i's *centroid maximin strategy*.

An equilibrium point s must always yield every player i *at least* his maximin payoff $\bar{u}_i$ because by (2.1) and (2.3) we have

$$U_i(s_i, s^i) \geqq U_i(\bar{s}_i, s^i) \geqq \bar{u}_i. \tag{2.4}$$

We call an equilibrium point s *profitable* to player i if $U_i(s) > \bar{u}_i$, and call it *unprofitable* to him if $U_i(s) = \bar{u}_i$. If s is profitable to *all* players it is called (uniformly) *profitable*. In the opposite case it is called (uniformly) *unprofitable*. If it is profitable to some players and unprofitable to others then it is called *semi-profitable*.

Likewise, a given game G as a whole will be called *unprofitable* to player i if it can be shown that no solution of G can yield player i more than his maximin payoff $\bar{u}_i$. Otherwise G will be called *profitable* to player i.

3

We shall assume that every game G will be preceded by a *bargaining game* $B(G)$, in which the players will try to agree on their payoffs and on strategies for obtaining these payoffs.[2] Only after the bargaining game has been completed will the players play the main game G itself, implementing the strategies agreed upon and obtaining the corresponding payoffs. We can analyze the bargaining game $B(G)$ by assuming that each player i will use a *bargaining strategy* d_i having the nature of a *decision rule* telling him whether to make a given *concession* (i.e., whether to accept a lower payoff than he has asked for so far) at any particular stage of this bargaining game or not. When we speak simply of "strategies," rather than "bargaining strategies," we shall always mean the strategies s_i of the main game G.

Suppose that at a given stage of the bargaining game $B(G)$ one of the players proposes some joint strategy $s \in S$ as the joint strategy to be used by the players in the main game G. Then the bargaining strategy d_i of each player i must specify whether he is to agree at this stage to the proposed joint strategy s or not. We shall denote by $D(d_i)$ the set of all joint strategies s agreeable to player i at a given stage of the bargaining game $B(G)$, as determined by this bargaining strategy d_i.

Our rationality postulates for game situations fall into two main classes: postulates of *rational behavior* (in a narrower sense) and postulates of *rational expectations*. The former essentially state the implications of the assumption that rational players prefer strategies yielding higher payoffs and are indifferent between strategies yielding equal payoffs. The latter state the implications of the assumption that each player will expect the other players, also, to act rationally, in accordance with their own real interests.

More specifically, our rationality postulates are as follows:

Rationality Postulates for Game Situations.
Class A: Postulates of Rational Behavior in a Narrower Sense.

Subclass A: Postulates of Preference for Strategies Yielding Higher Payoffs.*

A1. *Maximin Postulate.* In a game G *unprofitable* to you, always use a *maximin strategy* $\bar{s}_i$. (In other words, if you cannot hope to obtain *more* than your maximin payoff $\bar{u}_i$ anyhow, then use a strategy that will absolutely assure you at least *that much*.)

A2. *Best-Reply Postulate.* In a game G *profitable* to you, so far as your binding agreements with other players allow, always use a strategy s_i^* representing a *best reply* to the other players' strategy $(n-1)$-tuple

$(s^*)^i$. (This postulate implies that in a profitable non-cooperative game the players' strategy n-tuple s^* will always be an *equilibrium point*. For reasons we shall discuss below, the postulate does not apply to *unprofitable* non-cooperative games. In the case of *cooperative* games, the postulate does not limit the players' choice to equilibrium points because, as soon as the players *agree* on some joint strategy s not having the nature of an equilibrium point, the postulate ceases to be operative.)

A3. *Generalized-Best-Reply (or Expected-Utility Maximization) Postulate.* In the bargaining game $B(G)$ associated with any game G profitable to you, as far as your binding agreements with other players allow, always use a bargaining strategy d_i representing (at least) a *generalized* best reply to the bargaining strategies you expect the other players to follow. (In the bargaining game $B(G)$ the players do not know each other's bargaining strategies and so have to rely on the *subjective probabilities* they assign to various possible bargaining strategies d_j by the other players j. Therefore we cannot require more than that their bargaining strategies should be *generalized* best replies to the other players' expected bargaining strategies. In contrast, in the main game G itself we can require that the players' strategies should be *actual* best replies to one another (Postulate A2), because our theory will yield sufficiently specific predictions about the other players' strategies so as to enable each player to satisfy this stronger requirement in choosing his own strategy.)

A4. *Acceptance of Higher Payoffs.*

Part I. Let s and s^* be two joint strategies in set S, both of them consistent, with our other rationality postulates, but such that s^* would yield you (player i) a *higher payoff* $U_i(s^*) > U_i(s)$, Suppose that, at a given stage of the bargaining game $B(G)$, your bargaining strategy d_i would include s in the set $D(d_i)$ of joint strategies agreeable to you. Then your bargaining strategy d_i must also include s^* in set $D(d_i)$. (In other words, if you are willing to agree to some joint strategy s, you must be even more willing to agree to another joint strategy s^* yielding you a higher payoff than s would.)

Part II. Let $d = (d_1, \ldots, d_i, \ldots, d_n)$ and $d^* = (d_1^*, \ldots, d_i^*, \ldots, d_n^*)$ be two n-tuples of bargaining strategies for the n players, both d and d^* being consistent, with our other rationality postulates, but such that d^* would yield you (player i) a *higher payoff* u_i than d would. Suppose that you have started out using bargaining strategy d_i while the other players have started out using $d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_n$. Then you must be willing to accept, an agreement under which all players would shift to the bargaining strategies $d_1^*, \ldots, d_i^*, \ldots, d_n^*$.

Postulates A1 to A4 are essentially specialized forms of two rationality postulates commonly used in the theory of individual rational behavior

(decision theory): Postulates A2 and A3 being variants of the (Expected) Utility Maximization Principle,[3] while Postulates A1 and A4 being variants of the Sure-Thing Principle.

*Subclass A**. Postulate of Indifference Between Strategies Yielding Equal Payoffs.*

A5. *Equiprobability or Centroid Postulate.* Suppose that player i expects that strategies s_i, $s_i^*, \ldots$ would all yield him the same payoff u_i, and that all these strategies would be equally consistent with all our other rationality postulates. Then player i will be *equally likely* to use any of these strategies. Hence his behavior will be such as *if* he used the *centroid strategy* s_i^0 of the set S_i^0 consisting of all these strategies s_i, $s_i^*, \ldots$. (This postulate follows from the customary operational definition of *equality* between two utility payoffs. For instance, if player i were found to choose strategy s_i *more often* (i.e., with a *higher* probability) than strategy s_i^*, then this would have to be interpreted as an indication of his attaching a *higher utility* to it.)

Class B. Postulates of Rational Expectations.

B1. *Mutually Expected Rationality.* In the same way as you will yourself follow the present postulates if you are rational, you must expect, and act on the expectation, that *other* rational players will likewise follow these rationality postulates.

B2. *Symmetric Expectations.* You cannot choose your own bargaining strategy d_i on the expectation that a rational opponent will use a bargaining strategy d_j *more concessive* than you would use yourself in the same situation. (That is, if you would yourself in his place refuse a certain concession and would regard this refusal as rational behavior on your part, then you cannot, expect that another player, no less rational than yourself, will take a more accommodating attitude in that situation.)

B3. *Expected Independence of Irrelevant Variables.* You cannot expect a rational opponent to make his bargaining strategy d_j dependent on variables whose *relevance* for rational bargaining behavior *cannot be established* on the basis of the present rationality postulates. (The purpose of this postulate is to exclude some completely arbitrary decision rules, such as, e.g., making the players' payoffs proportional to their telephone numbers, etc. Many of these arbitrary decision rules would be quite consistent with all our other postulates. Our last postulate, however, rules them out on the ground that there is no reason to regard, e.g., telephone numbers as *relevant* variables in deciding the players' payoffs, etc.) More generally, the

postulate implies that the *only* variables influencing the players' bargaining behavior will be:

(i) the *payoffs* associated with alternative outcomes for each of the players, and
(ii) the *subjective probabilities* each player assigns to different possible outcomes being accepted or rejected by the other player(s).

Among these variables, only those mentioned under (i) are *independent* variables while the variables under (ii) are themselves determined by the variables under (i).

We have seen that our postulates of Class A (postulates of rational behavior) are closely related to the rationality postulates used in individual decision theory. Our postulates of Class B (postulates of rational expectations) have no direct counterparts among the latter. Nevertheless they represent a natural extension of Bayesian decision theory in the following sense.

Inherent in the use of subjective probabilities under the Bayesian approach is always the requirement that a rational individual must choose his subjective probabilities on the basis of the *best information* available to him, which may be called the Principle of Best Information. Our postulates of Class B are adaptations of this principle to game situations where each player is assumed to have completely reliable information about all other players' perfect rationality and intelligence. Our postulates of Class B require that, in assigning subjective probabilities to alternative possible strategy choices by the other players, each player should take full account of the fact that these other players are known to be highly rational and intelligent individuals.

We shall now try to indicate, as far as possible within the space available, how our rationality postulates can actually be used for defining rational behavior in specific game situations.

4

Playing a game effectively means solving the problem of choosing a rational strategy. It is convenient to subdivide this problem into several subproblems, which are in general not independent of one another, but are at least logically distinguishable:

1. *The enforcement or stability problem.* This consists in identifying the *stable* joint strategies, i.e., those which can be adopted by means of enforceable or self-enforcing agreements and therefore, once agreed upon, will in fact be implemented by the players.
2. *The joint-efficiency problem.* Let E be the set of all payoff vectors u that can be achieved by means of stable joint strategies. Then the joint-