1977

Toward A Theory Of Rational Interaction: Game Theory And Social Power

Andrew Kari Bjerring

TOWARD A THEORY OF RATIONAL INTERACTION: GAME THEORY AND SOCIAL POWER

by

Andrew K. Bjerring
Department of Philosophy

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario
July, 1977

ABSTRACT

The central problem explored in this dissertation is the nature of rational action in social, i.e. interdependent, contexts. The general thrust of its arguments is that to fully understand the nature of interdependence we must shift our focus from the overly restrictive a prioristic, static and individualistic approach to rationality of neo-classical micro-economic theory and game theory. Instead we must develop a dynamic model of rational interaction reflecting the way in which the actions of a rational agent and the social context of his behaviour (i.e. the actions of others) evolve jointly in an adaptive and rational way.

The two cornerstones of this exploration corresponding to these two senses of the term "rational" are the work of the economist John Harsanyi and the work of the social psychologists Thibaut and Kelly and Richard Emerson. The central bridge between the two approaches to rationality explored in this study is the notion of social power, first as it arises in Harsanyi's general bargaining theory, and secondly as it arises in exchange theory.

The stance adopted towards the game-theoretic treatment of interdependence is basically critical. The analysis begins with an examination of Harsanyi's recently

proposed theory of non-cooperative games and his so-called "tracing procedure". A novel interpretation of Harsanyi's procedure is used to draw out the assumptions lying behind his Bayesianism: In particular his provision of the Bayesian reasoner with prior probabilities regarding the behaviour of his "opponents", and the assumptions he makes regarding the reasoner's "motivations". It is then shown how the model fails in its avowed purpose of providing a formal structure admitting both a normative and a psychological, or positive, interpretation.

The second stage of the critique of traditional game theory begins with a discussion of the explanatory notion of social power. To set the stage for the examination of Harsanyi's contribution to this discussion a detailed critique of Goldman's recently published action-theory-based analysis of social power is presented. It is argued that Goldman's analysis is totally inadequate because of his rejection of a strategic foundation for his theory. That is, although "social power" is generally considered to be an explanatory notion, Goldman's failure to link it up with a normative model proves his undoing. This same fault is shown to arise in his action theory.

This critique leads to a discussion of Harsanyi's game-theoretic approach to power. It is argued that his n-person theory of cooperative games is an inadequate normative foundation for that concept because it fails to

make threats and coalition formation credible. Moreover, there are a number of unjustified assumptions he makes in developing the system of equations representing the equilibrium conditions among the set of coalition "agreements".

On these grounds his theory of social power is inadequate. That it also fails to provide the rational agent with any guide in exercising his power is taken as a secondary fault.

Finally, Emerson's exchange-theoretic approach to power-dependence relations is presented as a model of interaction in keeping with the second approach to rationality. It is argued that although exchange theory is an avowedly explanatory theory, it has a normative underpinning sufficient to support notions like social power, and thereby to provide a theory of rational interaction.

ACKNOWLEDGEMENTS

I would like to acknowledge the support and direction I have received during the preparation of this dissertation from my supervisor, Professor Jim Leach. During the five years of my apprenticeship he has been my intellectual guide, mentor and friend, and has contributed in innumerable ways to the development of my philosophical awareness and perspective, for which I thank him, but do not hold him responsible.

I would also like to acknowledge the assistance and guidance received from Professor Cliff Hooker, who commented extensively on an earlier draft of some of the chapters and with whom I have spent many happy hours discussing shared philosophical concerns.

Deep thanks go, too, to my wife Nancy for her companionship, understanding and her willingness to listen patiently to an endless succession of ideas in the making; and to my typist Susan Weekes for faultlessly completing a task no other sane person would have undertaken.

During the preparation of this thesis I was the grateful recipient of Canada Council Fellowships and a Queen Elizabeth-II Ontario Scholarship.

LIST OF FIGURES

LIST OF FIGURES

why the desires he has at that time are what they are. In this "procedural" sense of the term, then, magical ritualistic behavior, primitive religious behavior, and the behavior of modern economic man in the market place are equally "rational". Moreover, they are equally social, and present any general theory of behavior with the same perplexing puzzles.[1]

A theory of rational interaction, then, must confront the interdependence of men as well as their interacting. The problems besetting the theorist are enormous, of course. Not only must the whole of the social and behavioral sciences be sifted through for workable fingerholds, but the program to unify selected theories around the central model of social man will as often as not meet concerted opposition in the more empirically-minded sciences. This is why the enterprise is as much philosophical as it is scientific. What is being sought is a theoretical structure capable of more than experimental confirmation by a few selected phenomena. It must be capable of bridging the gaps among existing theories as well. It must unify, rather than replace. But it unifies around a normative model of man.

As mentioned, this dissertation hopes to provide a cornerstone for that larger enterprise. In moving toward a theory of rational interaction, it explores two general issues in the social sciences possessing normative and

ubiquitous. With only a few limiting exceptions, all of man's actions are rational, and we are to understand them in that light. However, as I will argue, there are two approaches to rationality in keeping with this general insight. But, at the general level, to attempt to reconcile these opposing views by suggesting that man is "sometimes" or "in some ways" rational, is not really to the point, for what is at issue is whether normative terms such as "rational" are to be conceived as giving man contact with the domain of the eternally true, or whether they are to be conceived as arising out of and as descriptive of, his experience. Thus conceived, the philosophical lineage of these common views touch upon every facet of formal philosophy. The history of very few philosophical issues is untouched by the contrast between them.

Not surprisingly, then, to venture an opinion on that fundamental clash is often to open the floodgates of criticism, for there are 2000 years of arguments on one side or the other ready to crush the unwary. Nevertheless, this dissertation does offer an opinion on this fundamental issue. More than that, it hopes to strengthen the case for the position that man <u>is</u> a rational animal, and that the explanatory understanding of his behavior had better incorporate an appreciation of his rationality or something is lacking.

In developing this thesis, I discuss two approaches

to rationality, both meant to defend this view. The first, the dual-interpretation model, sees the problem as being one of characterizing the term "rational", and then showing that man usually behaves in the appropriate way. The second, the procedural model, sees the problem as being one of characterizing the term "rational action", where the focus is already on man's behavior. According to this second approach, then, exactly how we are to understand the term "rational" is a function of how we understand man's actions. The two problems go hand in hand and amount to the development of a general theory of action. The problem thus viewed is not exhausted simply by characterizing the nature of rational choice in a few general and well-defined situations, as the dual-interpretation approach might have it, because part of the behavioral question concerns why the agent views the situation that way rather than another.

Consequently, and although there is appeal to a rough characterization along the lines of "behavior is rational to the extent that it is optimally goal-directed", our understanding of the terms "behavior", and "goal", are as problematic as our understanding of the term "optimal". Such a characterization, then, can offer no permanent solution.

Because of this sort of tension between the dual-interpretation and procedural approaches all that is being offered at this stage in defence of the thesis that man is a rational animal is a promissory note. If a general

theory of action incorporating the assumption of rationality can be developed and proves to be a deep, interesting and fruitful theoretical framework leading to both normative and explanatory insight, then an interesting dialogue with the essentialist can take place. Of course, if such a theory can be developed, then any strictly a priori defence of the thesis would have been irrelevant anyway.

This dissertation hopes to contribute to the development of that general theory of action. In particular, it confronts one central dimension of such a theory -- the nature of rational interaction. It pursues answers to the fundamental question "When man interacts with man, how are we to understand their behavior as rational?"

Note that this formulation of the problem does not equate the general theory with a theory of conflict management. That approach to rationality and to interaction is allied with the dual-interpretation model, and will be explored and rejected on the grounds that it denies important interdependencies of man upon man. That is, it denies the inherent sociality of man's behavior. What will be argued is that to understand an agent's behavior in terms of wants, values and beliefs is already to understand it as taking place in a social context, for those parameters are dynamic and change during and due to interactions with other agents. To understand an agent's behavior, then, is to understand why he views his situation the way he does, and

why the desires he has at that time are what they are. In this "procedural" sense of the term, then, magical ritualistic behavior, primitive religious behavior, and the behavior of modern economic man in the market place are equally "rational". Moreover, they are equally social, and present any general theory of behavior with the same perplexing puzzles.[1]

A theory of rational interaction, then, must confront the interdependence of men as well as their interacting. The problems besetting the theorist are enormous, of course. Not only must the whole of the social and behavioral sciences be sifted through for workable fingerholds, but the program to unify selected theories around the central model of social man will as often as not meet concerted opposition in the more empirically-minded sciences. This is why the enterprise is as much philosophical as it is scientific. What is being sought is a theoretical structure capable of more than experimental confirmation by a few selected phenomena. It must be capable of bridging the gaps among existing theories as well. It must unify, rather than replace. But it unifies around a normative model of man.

As mentioned, this dissertation hopes to provide a cornerstone for that larger enterprise. In moving toward a theory of rational interaction, it explores two general issues in the social sciences possessing normative and

explanatory faces. From recent attempts to provide a broader foundation for economic theory comes the theory of games. From virtually every branch of the social sciences comes the strategic but explanatory notion of social power. The exploration of recent approaches to these two issues is central to the development of the general theory, for what I hope to show is that in both cases present construals are inadequate, and what both problems require is an extended model of rational interaction in keeping with a procedural approach to rationality.

Thus, on the ashes of game theory conceived as a normative model with explanatory potential, and on the resolution of conflicting insights emodied within major extant approaches to social power as an explanatory notion with strategic overtones, I hope to justify the pursuit of an alternative approach to both. The hope is that this alternative will provide the beginnings of a conceptual framework around which the unification of the sciences of social man can begin.

1.2. Games

Neo-classical economics has clearly been the discipline most associated with the attempt to develop an explanatory theory on a normative foundation.[2] Of particular concern in this study are the theoretical progeny of one classic exploration of such a foundation, John von-Neumann's and Oscar Morgenstern's Theory of Games and Economic

Behavior.[3] More particularly still, I will be concerned with the theories proposed by the economist John Harsanyi, perhaps the most eminent game theorist of the past twenty years. Harsanyi is also one of the most technically-competent theorists to have recognized and explored the potential game theory possesses as a unifying framework in the social sciences. Indeed, though his concern is rarely with applications per se, he appears to be as concerned with showing why rational people behave in accordance with his theoretical results as he is in the mathematics of game theory. In short, Harsanyi has been chosen because he adopts the dual-interpretation approach to rationality mentioned above. His primary concern is with developing a normative model, but he is also committed to showing that it has a psychological, or positive interpretation.[4]

Informally, a game may be conceived as a "context of strategic interaction" among a set of agents. Each has available alternative choices of action, the consequences of which are a joint function of the choices of all the agents. Parlor games such as tic-tac-toe, checkers, and chess are classic examples of games in this sense, although the connotations those pastimes have of "play" and "enjoyable" are not to be transferred to the more technical notion.

Moreover, there are more general sorts of interactions also classifiable as "games". Agents bargaining over the selling price for a house, runners choosing racing strategies,

foundation of game theory cannot generate the explanatory notion of power. I argue that this failure is attributable to the adoption of the dual-interpretation approach to rationality.

I next turn to the third major alternative, the interactionist notion that social power is not simply a matter of one agent's abilities, but is also a function of the other agent's attitudes towards what the first can do. This approach takes power to be a function of the ability of an agent to control outcomes for another and of the alternatives there are available to the agent. Richard Emerson's notion of power-dependence relations is presented as an instance of this approach. But more importantly, it is his theory I plan to explore as a possible basis for a general theory of rational interaction.

The discussion of social power then can be viewed as a bridge between the critique of game theory and the dual-interpretation model of rationality and the presentation of an alternative framework for rational interaction of the procedural sort. On the one hand it is a case study highlighting the narrowness of game-theoretic modelling of social interaction. On the other it is a case study exemplifying the depth of the exchange-theoretic approach. In any case, it is a key theoretical notion in the social sciences, and a sufficiently general treatment will provide part of the means for achieving a unification of the sciences of man.

players will always choose a strategy such that the expected "payoff" to each agent is a specifiable value. They have thus offered a "solution concept" for all two-person constant-sum games.[6]

Most real interactions, of course, cannot be modelled as having this "constant-sum" characteristic. That is, most choice situations present options to the agents which lead to mutually advantageous or mutually disastrous results. These are called "variable-sum", or "mixed-motive" games, and present the main challenge to those whose primary interest is the explanatory use of game theory. Interactions of this sort will be focused on here. More specifically, I explore Harsanyi's theory of games as it pertains to two types of mixed-motive games. First, in order to clarify the dual-interpretation model, I explore his so-called "non-cooperative" theory, which treats situations where the agents cannot communicate and must choose their strategies independently. Secondly, to see how this approach deals with explanatory notions, I explore his "cooperative" theory which treats situations where full communication is allowed and jointly-chosen strategies are assumed to be admissible and fully enforceable once an agreement is reached.

I want to argue that whatever the inherent interest of game theory, and whatever its ability to model certain forms of interaction, it is not suitable as the sought-for foundation for a theory of rational interaction. Of

particular concern will be its assumptions pertaining to what each agent knows about the situation and the other agents, and the limitation of the game form to being a static reconstructive device. These arguments and others will occupy our attention as I try to show that although game theory cannot form the foundation for our theory, the reasons why that is the case point to a ready alternative.

1.3. Social Power

The second central theme of this study is the controversy surrounding the explanatory notion of "social power". The main issue seems to be how the relationship between agents A and B is to be understood when A is said to have "power" over B. I want to argue that although the notion is explanatory, it has normative elements which must be elaborated in a theory following the second, more general, approach to rationality -- the procedural approach.

It has become traditional in the literature on social power to distinguish between various "dimensions" of that notion, and these distinctions will be maintained here.[7] For example, we want to avoid confusing the amount of power an agent possesses and that he is capable of <u>exercising</u> during a given interaction. We also want to distinguish between that by virtue of which he possesses power, and the <u>amount</u> of power possessed. The former is usually called the base of the agent's power. For example, John Rockefeller may be very powerful because

he is very wealthy, but the President of the United States is possibly just as powerful for totally different reasons. The amounts of their power may in some general sense be "equal", but the bases are totally different.

This example also raises the distinction between so-called "legitimate" power, or "authority", and other sorts. The former is power possessed by an individual by virtue of his occupying a certain position in a formal social hierarchy. Presidents, priests, princes, judges, policemen, and army generals all possess "legitimate" power in this sense, although none may be either wealthy or physically powerful. However, although the examination of the mechanism of legitimation is an important problem in the social sciences, it will not be pursued here. It is the effects of power that are of primary concern, so I will treat legitimation as simply another power base, however unique and central to other discussions.

But the effects of social power are as varied as are the theories themselves. The ability to affect the state of the world, the ability to influence another agent's behavior (or the probability of his performing some specific act), the ability to affect another agent's "outcomes", and the ability to create a net "force" in a region of another's "life space" have all received attention in recent years as the focus for the effects of social power.

The last of these focal points is clearly of special interest to field theory, and so will not concern us here.

The task of linking field theory with other psychological theories will be left for another day.[8]

The others can be looked upon as providing the three main alternative approaches to power. Each will be explored. The first represents a theory without a strategic foundation; the second is in keeping with the dual-interpretation model; the third, I will argue, adopts a procedural approach to rationality.[9]

The first, power over the state of the world, is the line adopted by Alvin Goldman. His concern is to formulate a notion of social power in keeping with his approach to action theory. Power is to be related to "ability" in the sense that <u>given the way the world is</u>, the "ability" to enact certain behaviors entails the possession of power and vice versa. His analysis of <u>social</u> power, then, is an attempt to extend this foundation to include costs and interaction effects. I argue that this attempt fails. What Goldman's model most clearly lacks is a sense of the strategic dimension to social power. An agent has social power not only when and because he is in a position to manipulate, control and influence the world, but also when and because he is <u>believed by others</u> to be in that position. For example, a brutish-looking man with a reputation for having a terrible temper will be able to influence others even if he is, behind the reputation, as meek as a lamb. On the other hand, actually having such a temper may be worthless in a given situation if the opposite number is

ignorant of the fact. Such an "other" does not know enough to be afraid of the consequences of conflict, and his ignorance may give him an advantage lacked by a more knowledgeable substitute.

It is for reasons such as this that the second major approach alluded to above might be thought to be preferable. Treating power as the ability to influence the behavior of others at least leaves it open whether that power is based in great technological control over the state of the world, or is based in a set of beliefs shared by the interacting parties. This approach is followed by John Harsanyi, among others, who builds a model of social power on a game-theoretic analysis of the relative costs of conflict to the opposing agents.

In exploring Harsanyi's theory, however, I argue that game theory is unable to provide the strategic foundation needed for a notion of social power because, among other things, it cannot make "threats" credible to rational agents. Game theory assumes at the outset that the agents are minimally rational in the sense that they will at least not do themselves a disservice relative to an agreement already in hand. But, then it must also assume that they will not actually carry through any threat strategy intended to "scare" an opponent into more agreeable terms. No <u>additional</u> postulates of rationality can alter this basic assumption that rational players, according to game theory, <u>know</u> that threats will not be played. Thus, the normative

FOOTNOTES

1. That this position is not fully appreciated in the social sciences is clear from a study of the essays in Bryan Wilson's Rationality (Harper, 1970). One of the fundamental clashes arising out of the anthropological topics explored there is between those who construe primitive people's ritualistic modes of behavior as somehow at odds with instrumentalistic analysis, and those who seek a theory which sees a degree of social ritual in all behavior, and a degree of instrumentalism in all ritual. The latter point of view is entailed by the "procedural" approach to rationality, and is endorsed here.

2. For a discussion of the dual function of rationality in such theories see J. Leach, "The Dual Function of Rationality" (1976). The notion of the "dual-interpretation" approach to rationality is also discussed there.

3. John von-Neumann, and Oscar Morgenstern, 1944 and 1947.

4. The major works of Harsanyi in game theory are listed in the bibliography. Of particular concern in Chapters 2-4 are his 1975 papers, while Chapter 6 will focus on his 1959, 1962 and 1963 papers.

5. For an introductory discussion of utility theory in the context of the theory of games, see Luce and Raiffa, Games and Decisions (1957), especially Chapter 2.

6. Ibid., Chapter 4.

7. For a general discussion of these distinctions see Robert Dahl's, "The Concept of Power" (1957), and John Schopler's, "Social Power" (1965).

8. For a discussion of the field-theoretic approach see Cartwright and Zander's, "Power and Influence in Groups" (1968), and French and Raven's, "The Bases of Social Power" (1968).

9. Goldman's main contribution is discussed in Chapter 5, Harsanyi's in Chapter 6, and Emerson's in Chapter 7.

10. In Hooker, Leach and McLennen, Foundations and Applications of Decision Theory (Reidel, forthcoming).

1.4. Interaction and Social Power

The third and final focal point of this study, then, is the presentation of what I take to be the most serious contender as a general theoretic-framework providing a model of rational interaction. It is not offered as a replacement for game theory in what game theory sets out to do, since it is not a normative theory of the dual-interpretation sort. Instead, being an explanatory theory its normative foundation, if it has one, must be of the procedural sort. What I argue is that, first of all, Emerson's exchange theory has a normative foundation. Secondly, I argue that the sense of "rationality" it offers construes social interaction as rational at the very outset. The way in which man interacts is, simply, a "rational" way, and it is this sense of a natural social process being "rational" that gives the label "procedural" to this general approach.

Of course, the interpretation of Emerson's theory as offering a normative model encounters important philosophical obstacles. First, his theory is built around a theoretical primitive referring to social-structure. The notion of an "exchange relation" is not eliminable from his theory. Similarly, the property of "dependence" is a relational property attributable to the exchange relation. How we should construe such a theory as offering a normative account of behavior is at first glance problematic.

Secondly, it is clear that the "procedural" approach to rationality imports empirical content into the normative theory. It cannot help but do so since, in effect, the theory it offers is explanatory in the first place.

Thus, the adoption of the procedural approach to rationality raises the twin spectres of a priorism and individualism. Both must be fended off, if not defeated, if the procedural approach is to be justified. I attempt this in the following way.

First, I argue the plausibility of social-structural primitives in explanatory theories by showing how the usual psychological-reductionist arguments are mistaken. Next, I argue that normative theories must import explanatory constructs since the terms of appraisal they define always refer to theoretically construed phenomena. In this case, "rationality" refers to "interdependent behavior". But the latter notion is clearly dependent on how the best explanatory theories construe behavior and interdependence. Finally, I conclude that the plausibility of the need for social-structural primitives in explanatory theories regarding social behavior, therefore justifies their incorporation in the normative theory as well. The result is a prima facie case for the development of a normative theory built on an explanatory-theoretical-framework incorporating a social-structural primitive.

1.5. Plan of the Dissertation

The sequence in which these issues are discussed is as follows. In Chapters II through IV, I critically examine Harsanyi's theory of non-cooperative games based on the so-called "tracing procedure". This section draws from my previously published "The Tracing Procedure and a Theory of Rational Interaction".[10] The objective is to explore the dual-interpretation approach to rationality, as well as to criticize the specifics of Harsanyi's models.

Next, in Chapter V, I examine Goldman's theory of social power as one part of his attempt to generate a general theory of action. The lack of a normative foundation for his general theory is shown to be a crucial failing when it comes to providing a foundation for the notion of social power. Harsanyi's theory of cooperative games and social power is then discussed in Chapter VI, and I offer what I take to be sufficient reasons for its rejection. Although the normative foundation is clearly present, the dual-interpretation approach cannot support the development of explanatory notions like social power, which are at one and the same time linked with normative considerations.

Two bridges between this discussion and Emerson's exchange theory are then constructed. First, in Chapter VII, I impose some additional structure on the conflict among the three theories of social power, and opt for Emerson's approach. Secondly, in Chapter VIII, I generate general

criticisms of game theory and counter some *prima facie* objections to the adoption of exchange theory as a normative model.

In Chapter IX, I present more detail on how I see exchange theory operating as a normative foundation for a dynamic theory of rational interaction. Extension of the model thus generated into other areas in the social sciences is left for future work.

FOOTNOTES

1. That this position is not fully appreciated in the social sciences is clear from a study of the essays in Bryan Wilson's Rationality (Harper, 1970). One of the fundamental clashes arising out of the anthropological topics explored there is between those who construe primitive people's ritualistic modes of behavior as somehow at odds with instrumentalistic analysis, and those who seek a theory which sees a degree of social ritual in all behavior, and a degree of instrumentalism in all ritual. The latter point of view is entailed by the "procedural" approach to rationality, and is endorsed here.

2. For a discussion of the dual function of rationality in such theories see J. Leach, "The Dual Function of Rationality" (1976). The notion of the "dual-interpretation" approach to rationality is also discussed there.

3. John von-Neumann, and Oscar Morgenstern, 1944 and 1947.

4. The major works of Harsanyi in game theory are listed in the bibliography. Of particular concern in Chapters 2-4 are his 1975 papers, while Chapter 6 will focus on his 1959, 1962 and 1963 papers.

5. For an introductory discussion of utility theory in the context of the theory of games, see Luce and Raiffa, Games and Decisions (1957), especially Chapter 2.

6. Ibid., Chapter 4.

7. For a general discussion of these distinctions see Robert Dahl's, "The Concept of Power" (1957), and John Schopler's, "Social Power" (1965).

8. For a discussion of the field-theoretic approach see Cartwright and Zander's, "Power and Influence in Groups" (1968), and French and Raven's, "The Bases of Social Power" (1968).

9. Goldman's main contribution is discussed in Chapter 5, Harsanyi's in Chapter 6, and Emerson's in Chapter 7.

10. In Hooker, Leach and McLennen, Foundations and Applications of Decision Theory (Reidel, forthcoming).

CHAPTER II

INTRODUCTION TO HARSANYI'S TRACING PROCEDURE

2.1. Introduction

The central issue confronting any theory of rational interaction is the problem of interdependent choice. Put simply, the key element of that problem is how rational people behave in their dealings with one another when what happens to them is a joint function of their individual behaviors. Game theory, as one theory of rational interaction, confronts this issue head on.[1]

The first step in traditional game theory's treatment of this problem is the decision to deal with one particular "form" of interaction as paradigmatic of all interaction contexts of rational individuals. This so-called "normal form of representation" has proven to be very powerful insofar as it is capable of generating intriguing technical puzzles and seems to have interesting mathematical properties. Of course, most game theorists claim more on behalf of the normal form than formal elegance, for it is advanced as a model for rational interaction.[2] The assumptions it embodies regarding the notion of a rational agent, however, are very restrictive.

The normal form is partially defined by the assumptions which specify the structure of any context of interaction

between rational agents. These can be reduced to the following four:

1) There is a fixed set of "players" whose actions affect the outcomes and each of whose outcome is in turn affected by some actions of the others.

2) Each player has a fixed set of strategy options among which one and only one may be chosen, that choice to be made at some fixed time.[3]

3) There is a fixed pattern of consequences of possible strategic interactions among the strategies available to the players.

4) There is a fixed utility assignment to each consequence by each player.

On the other hand, there are equally important epistemic assumptions which help interpret the normative meaning of the normal form:[4]

1) All players are known to be rational by all players, and all players know this.

2) All elements of the interaction structure itemized above are known by all players, and all players know this.

The first of these epistemic assumptions is usually called, for obvious reasons, the "assumption of mutual rationality". When conjoined with the second assumption and the structural definition it gives rise to an even stronger assumption of much the same form called the "transparency of reasoning assumption". In this latter form the axioms generating the normal form confront game theory with its central puzzle.

Consider a two-person game. If everything that is relevant to one player's decision-making is also known by

his "opponent" and both are rational in the very same way, then the result of each player's reasoning including the choice he is about to make is public knowledge. But the result of an opponent's reasoning is, one might expect from the interdependency, either information which a rational player requires before he starts his own reasoning, or else it is information which, once "received", will in general drive him into a new round of reasoning incorporating this new information in some way. Now the former is clearly not the case since if waiting for the results of an opponent's reasoning were rational, then in general all players in a game of the normal form would be forever waiting. Therefore, we must assume that it is rational to take the results of an opponent's reasoning into account in an iterative calculation of some sort. However, the conclusion of any round of reasoning by any player will merely start his opponent off on another cycle, and we are into the familiar "I think/he thinks" reciprocal-reasoning regress.

Of course some sorts of "normal games" are such that this reciprocal-reasoning regress does not get vicious. In general this will be the case whenever the game is such that the players' choices at some iteration in the regress just so happen to "mesh" with one another in such a way that their being made public does not alter anyone's decision. Two-person zero-sum games, for example, are "solved" in this way by the trick of each player "rationally" choosing a mixed

strategy of just that sort that his opponent is unable to counter with any alternative strategy which will do anything to improve his payoff.[5] On the other hand, some variable-sum games just happen to have "obvious" strategy choices which are in some sense to the "mutual advantage" of all players, and so have the important property that were each player to announce his intention none of his opponents would wish to alter his choice of strategy. Such obvious equilibrium points are clearly paradigms for the sorts of solutions which game theory is committed to devising.

In general, however, the transparency of reasoning assumption and the restrictive normal form of representation have together confronted game theory with an apparently insoluble problem. Either we show that this reciprocal reasoning regress will somehow end for all games or else we must come up with a whole bagful of "tricks" of the sort introduced for two-person zero-sum games. Since neither of these approaches has so far been successful, traditional game theory has up to now appeared as a patchy collection of partial answers, and the normal form has been a technically intriguing but philosophically frustrating foundation for a normative theory of interaction. Of course, the option is always available of dropping one or both of the constraints on the normal form and developing another paradigm for rational interaction. As it happens, this course is being followed by an increasing number of theorists in recent

years.[6] The major exception to this trend, however, is perhaps the most eminent game theorist of the past two decades, John Harsanyi.[7]

In recent papers[8] Harsanyi has proposed a very sophisticated reasoning algorithm which he claims defines a "solution" for all n-person non-cooperative games. Taken together with his work in cooperative game theory (see Chapter 6), what he is in effect offering us is a complete Bayesian theory of games.

Anyone who takes the quest for a theory of rational interaction seriously cannot fail to be impressed by Harsanyi's program and the importance of his work were it to succeed. On the other hand, it is my firm belief that this most recent proposal of Harsanyi's places game theory squarely at the crossroads. Should it, too, prove to be a failure we would, I think, be well advised to abandon traditional game theory and begin to look much more seriously at the alternatives which are beginning to surface.

It is one of the theses of this dissertation that, regretably, Harsanyi's program has failed and with it should be interred traditional game theory, the normal form of representation,[9] and the assumption of the transparency of reasoning. This general thesis will be discussed in Chapter VIII.

2.2. The Tracing Procedure and Bayesian Game Theory

It has been argued above that for traditional game theory the problem of interdependent choice takes the form of the reciprocal-reasoning regress. As might be expected, the usual Bayesian analysis of this regress (hereafter called the "naive" Bayesian approach[10]) is to assume that at the outset the players each assign subjective prior probabilities to the strategy options of their opponents. The regress proceeds with repeated calculations of "best replies", that is, those strategy responses which, given the opponents' choices, will maximize that player's expected payoff.

Unfortunately for naive Bayesians, two problems prevent this approach from succeeding. First, unless the prior probabilities on which each player begins his analysis of the game are objective, the results of even the first best-reply-calculation cannot be public knowledge. This does not mean that the various players' reasonings cannot converge to the same point, but it does make the job of showing not just that they do, but that they will always converge next to impossible, and given the approach it is "convergence" which allows each player to stop the regress. Secondly, even if objective priors could be defined, for most games and many prior probability assignments this form of iterative reasoning does not converge to an equilibrium set of strategies anyway, and so cannot define a general solution concept.

A game Γ can be characterized by a vector

$$U = (u_1^1, \ldots u_1^K; u_2^1, \ldots, u_2^K; \ldots; u_n^1, \ldots, u_N^K)$$

whose components are the payoffs to each of the n players for each of the K pure strategy n-tuples. (This vector is equivalent to the matrix normal form.)

Letting all games of a given size be represented by the symbol $\mathcal{G}(n; K_1, \ldots, K_n)$, we define the term <u>almost all</u> in the following way. We say the statement ∂ is true for <u>almost all</u> games if for all $\mathcal{G}$, the set of $\Gamma \varepsilon \mathcal{G}$ for which ∂ <u>is false</u> is a set of measure zero relative to the dimensionality of $\mathcal{G}$.

Finally, a strategy s_i^* is a <u>best reply</u> for player i to a given (n-1)-tuple $\bar{s}_i$ if:

$$u_i(s_i^*, \bar{s}_i) \geq u_i(r_i, \bar{s}_i) \text{ for all } r_i \varepsilon S_i.$$

with a viable resolution of the regress.[12]

Before proceeding to a more detailed discussion of Harsanyi's work a brief word on "interpretations" is in order. The normal form of representation is in effect a formal mathematical structure requiring some sort of semantical "interpretation", or model. The usual model is the normative interpretation from game theory: the players are taken to be ideally rational agents, and the context of interaction is purely hypothetical, i.e. is not assumed to model any actual context of interaction. Harsanyi, however, is not content with this very sterile conception of the meaning of the theorems of game theory. Although he recognizes that the normal form was generated with that sort of model in mind he is equally concerned with generating a psychological, or positive interpretation for the normal form and the theorems he proves about it. In other words, the steps in his tracing procedure, as with the other solution concepts he generates, must be acceptable as a model of a least some actual human reasoning processes, or else he will view the result as inadequate. It is this concern that leads to the labelling of his approach as the "dual-interpretation" model. I will be concerned with both aspects of his theory, but the central thrust of the critique will be that his game theory cannot function as the foundation for a general theory of rational interaction.

2.3. Preliminary Definitions

Before proceeding to a discussion of the tracing procedure there are a number of symbols which must be defined. With the exception of a few simplifications, the symbolism is Harsanyi's. However, I will duplicate only those of his definitions which our purposes demand.

n = number of players

i = the label for the ith player (i.e. $i = 1,2...n$)

K_i = the number of pure strategies available to the ith player

a_i^k = the kth pure strategy of player i

$K = \prod_{i=1}^{n} K_i$ = the number of possible pure strategy n-tuples

b^k = the kth pure strategy n-tuple (the order is arbitrary) $1 \leq k \leq K$

s_i^k = the probability of player i playing a_i^k $(1 \leq k \leq K_i)$ in a given mixed strategy

$s_i = (s_i^1, s_i^2, \ldots, s_i^{K_i})$ = the (mixed strategy) probability vector for player i (i.e. all s_i^k are ≥ 0 and ≤ 1 and $\sum_{k=1}^{K_i} s_i^k = 1$)

$S_i = \{s_i : \forall_k \ (1 \leq k \leq K_i \supset s_i^k \geq 0) \quad (\sum_{k=1}^{K_i} s_i^k = 1)\}$

= the ith player's strategy space (a simplex of $K_i - 1$ dimensions)

$S = S_1 \times S_2 \times \ldots \times S_i \times \ldots \times S_n$ = the collective strategy space (a convex and compact polyhedron of $(\sum_{i=1}^{n} K_i)$ - n dimensions)

$\bar{s}_i = (s_1, s_2, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ = the collective strategy opposing player i

$\bar{S}_i = S_1 \times S_2 \times \ldots \times S_{i-1} \times S_{i+1} \times \ldots \times S_n$ = the collective strategy space opposing player i

Π_i = player i's subjective ("estimated") probability distribution across the (n-1) other players' (K/K_i) possible pure strategy (n-1)-tuples

Π = the "estimated" probability distribution across all n players K possible pure strategy n-tuples (Π is a function of the parameter 't', see below)

p = the prior probability distribution across the K possible pure strategy n-tuples

$\bar{p}_i$ = the prior probability distribution across the (K/K_i) possible pure strategy (n-1)-tuples of pure strategy combinations of the players opposing player i

$Q_i(s_i) = \{a_i^k : s_i^k > 0\}$ = the carrier of s

$Q(s) = \bigcup_{i=1}^{n} Q_i(s_i)$ = the carrier of the n-tuple s

u_i^k the payoff to player i when the players select the pure strategy n-tuple b^k

$U_j(s)$ = payoff to player j when each player i $(1 \leq i \leq n)$ plays s_i

t = a parameter $(0 \leq t \leq 1)$ with the pre-systematic character of "time into the solution (i.e. 'reasoning') procedure"

A game Γ can be characterized by a vector

$$U = (u_1^1, \ldots u_1^K; u_2^1, \ldots, u_2^K; \ldots; u_n^1, \ldots, u_N^K)$$

whose components are the payoffs to each of the n players for each of the K pure strategy n-tuples. (This vector is equivalent to the matrix normal form.)

Letting all games of a given size be represented by the symbol $\mathcal{G}(n; K_1, \ldots, K_n)$, we define the term <u>almost all</u> in the following way. We say the statement ∂ is true for <u>almost all</u> games if for all $\mathcal{G}$, the set of $\Gamma \varepsilon \mathcal{G}$ for which ∂ <u>is false</u> is a set of measure zero relative to the dimensionality of $\mathcal{G}$.

Finally, a strategy s_i^* is a <u>best reply</u> for player i to a given (n-1)-tuple $\bar{s}_i$ if:

$$u_i(s_i^*, \bar{s}_i) \geq u_i(r_i, \bar{s}_i) \text{ for all } r_i \varepsilon S_i.$$

FOOTNOTES

1. Throughout this essay I want to distinguish between "game theory" proper, and the class of theories whose subject matter is the interaction of rational agents. Included in the latter sort of theory would by dynamical considerations and interdependencies untouched by traditional game theory. It is one of the theses of this dissertation that the problem of interdependent choice must be posed within the context of the more general approach to rational behavior. The approach to rationality it embodies, of course, is the "procedural" rather than the "dual-interpretation".

2. A possible exception to this generalization is Anatol Rapoport who has long maintained a rather unique position with regard to game theory. He views the theory itself as a branch of mathematics, and takes it to be stimulating and uncontroversial. On the other hand, the prescriptive use of game theory is, he thinks, unwarranted, and in need of drastic revision. (See, for example, his (1966), (1970), and his contributions in his (1974).)

3. The very notion of "strategy" is, of course, very specifically defined. If construed in its strict sense it entails that a full "life plan" is being selected at every choice junction. Of course, in any particular case an agent's options may differ only in very restricted ways, and so we may safely characterize each strategy simply by the ways in which it differs from its alternatives.

4. Assumptions regarding "states of knowledge" are not, strictly speaking, either "formal" or "normative", but rather, border on the "psychological". However, the usual way of expressing the normative interpretation of the normal form is in terms of the "rational agent" and what he knows, believes and values. This is the standard albeit regretably sloppy way of speaking about the axioms of utility and probability theory. I will continue this practice, and use this "semi-psychological" model as representing the normative interpretation of game theory. The reader should note, however, that the so-called "rational agent" is not subject to any broader sorts of psychological

considerations than those covered by the axioms of utility theory, probability theory, and game theory, and that the provision of a positive interpretation for the model must involve more than these rudimentary descriptions.

5. In general, selecting a randomized strategy is one way theories based on the normal form have of blocking information from reaching an opponent. However, this blockage is purchased at the price of admitting there is no purely "rational" choice among the pure strategy options.

6. Foremost among these mavericks has been Thomas Schelling, who was one of the first theorists to call for a "reorientation" of game theory. (See his The Strategy of Conflict (1960).) Nigel Howard also falls into this camp, with his "meta"-game theory offering a provocative alternative to the normal form. (See his The Paradoxes of Rationality (1971).) David Gauthier also fits this description. (See his "Reason and Maximization" (1975).)

7. Harsanyi's contribution to the theory of games is unquestionable. Many, though by no means all, of his important papers in this field are listed in the Bibliography. In particular, see his (1961), his (1966b), and Chapter VI of this dissertation.

8. See Harsanyi's papers on the Tracing Procedure (1975a), (1975b), and (1975c). Copies of (1975a) and (1975b) were distributed at the Fifth International Congress of Logic, Methodology and Philosophy of Science held at The University of Western Ontario, London, Ontario, August, 1975. The content of these two papers has also been published in (1975c). Page references will be to the earlier papers.

9. The "formalist approach" will often refer to this dependence of game theory on the normal form of representation. More generally, I take that term to cover any approach to normative matters having the following two properties: First, it accepts a fixed, formal, axiomatic structure as its theoretic foundation. Secondly, it contends that normative notions are derived from this purely structural (non-empirical) framework by way of a semantical interpretation. Thus, within utility theory the usual interpretation of the axioms is such that the normatively "rational" choice is that which maximizes expected utility. "Formalization" per se is, of course, not the problem so much as

the way in which the formal structure is generated and altered. Insight into normative issues comes less from fiddling around with axioms than from studying the sciences of man, and it is this that the formalists have forgotten. More particularly, the "procedural" approach to rationality is clearly at odds with the formalist approach, as the insights it embodies must come from explanatory theories.

10. Of course, no one who calls himself a "Bayesian" ever advances this naive analysis. What I intend by this label is simply that the analysis is in keeping with the fundamental Bayesian approach in decision theory.

11. What Harsanyi has told us, however, is sufficient to permit a pretty good guess as to what he has in mind. See Chapter IV, section 4.2.3.3.

12. The notion of a "filter" is borrowed from information theory where its meaning is roughly "any device whose input and output can be considered to be 'information', and where the output at a given time is some function of past inputs".

CHAPTER III

HARSANYI'S TRACING PROCEDURE

3.1. Introduction

This chapter examines two alternative ways of presenting Harsanyi's tracing procedure for both his linear and logarithmic models. One is Harsanyi's, the other is mine. I then proceed to a detailed discussion of how these procedures deal with a very simple form of two-person game. In Chapter IV, I undertake a critique of the approach.

3.2. The Linear Tracing Procedure (LTP)

3.2.1. Non-Technical Overview

Harsanyi's linear tracing procedure is an algorithm which implicitly defines a solution concept for "almost all" n-person non-cooperative games. The essence of the procedure is the "tracing-out" of a path in an algebraic space based on the game of concern, Γ, from a point uniquely determined by the specification of the prior probability distribution, p, to a point in the space corresponding to an equilibrium point of Γ. When the equilibrium point selected out by the tracing is unique, which is the case for almost all games and almost all prior probability distributions,

it is the "solution" for the game.

There are two equivalent ways of representing the actual tracing procedure; they differ primarily in the definition of the algebraic space within which the "tracing" takes place. Harsanyi's version (Version I) defines a space based on an infinite family of games $\{\Gamma^t\}$, $0 \leq t \leq 1$, each closely related to the game of concern (Γ). The two most important characteristics of this family of games are, first, that any equilibrium point of Γ^0 is a collective best reply to the prior probability distribution, and secondly, that Γ^1 is identical to Γ, and so all equilibrium points of Γ^1 are equilibrium points of Γ. In this version the tracing occurs in a space defined around the equilibrium points of the games $\{\Gamma^t\}$. Any particular "trace" follows the path connecting the point corresponding to the equilibrium point of Γ^0 (given the prior) through intermediate points corresponding to the equilibrium points for the intermediate members of the family of games, through to a point corresponding to the equilibrium point for the game Γ^1, that is, Γ. The equilibrium point corresponding to the end point of the traced path is then dubbed the "solution" of the game.

The second version of the linear tracing procedure (Version II) is only hinted at by Harsanyi. I have developed it in detail in order to emphasize its connection with the naive Bayesian reasoning model. It traces out pathways in the actual collective strategy space for the game rather

sets. Such a branching α may be conveniently described by the symbol:

$$(3.3) \quad \alpha = \{\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_{\ell-1}, \begin{matrix} \bar{\alpha}^1_\ell \\ \bar{\alpha}^2_\ell \end{matrix}, \begin{matrix} \bar{\alpha}^1_{\ell+1}, \ldots, \alpha^1_{E_1} \\ \bar{\alpha}^2_{\ell+1}, \ldots, \alpha^2_{E_2} \end{matrix}\}$$

where $\alpha^1_{E_1}$ and $\alpha^1_{E_2}$ are different equilibrium points of the game.

3.2.4. A Final Word on Version II

There is another revealing way of viewing this interpretation of the LTP. The players may be looked upon as a mountain climbing team with a goal of "climbing" a rather convoluted terrain to the highest point possible given their starting point and governed by the following two constraints. First, they must take only infinitesimal "steps". Second, those steps must be in the direction of maximal slope upwards. To decide which direction that is, each player must compute the direction in the terrain of his "best reply" (the maximal slope as he sees it), and the team as a whole non-cooperatively combines these "suggestions" into a collective decision. For most steps a player will "suggest" the same direction as at the previous step, and so the team as a whole proceeds in a straight line. However, at some points in the climb (i.e. switching points), a player will alter his estimate of the best direction and therefore the team modifies its assessment of the optimal direction of ascent.

Given the details of the LTP (which in effect defines

3.2.2. LTP Version I

Harsanyi's presentation of the linear tracing procedure as applied to a game Γ with payoff vector $U=(u_1,u_2,\ldots,u_n)$ begins by considering a one-parameter family of games $\{\Gamma^t\}$, $0\leq t\leq 1$, where each game Γ^t has the following characteristics:

1) For every game Γ^t each player i has the same strategy space S_i as in Γ.

2) The payoff function for each player i in game Γ^t is given by:

$$(3.1)\quad v_i(s_i,\bar{s}_i;p,t) = t\,u_i(s_i,\bar{s}_i) + (1-t)\,u_i(s_i,\bar{p}_i)$$

As discussed above, $\Gamma^1 = \Gamma$, while Γ^0 is a game in which each player's payoff is a function only of his choice of strategy and the prior probability distribution, not the other players' strategies. The set of "best replies" in Γ^0 is, therefore, an equilibrium point in Γ^0 and happens to coincide with the set of best replies in Γ to the priors, p. In Γ, however, these particular strategies would not usually be in equilibrium.

Now, letting E^t be the set of equilibrium points in game Γ^t (for a given prior p), Harsanyi shows that the set of points $P = \{x: x=(t,s) \wedge t\in[0,1] \wedge s\in E^t\}$ is almost always a family of one-dimensional, piece-wise algebraic curves. The linear tracing procedure consists in following these "paths" in P, starting at $(0,s^*)$ where s^* is the equilibrium point of Γ^0. (It is almost always unique for a given p.)

Such paths in P always exist and (for a given p) almost always lead to a unique point $(1,s^{**})$. The point s^{**} is always an equilibrium point of Γ and, where the point $(1,s^{**})$ is unique, s^{**} may be called the "solution" of the game Γ.

3.2.3. LTP Version II

The interpretation of the LTP to be outlined in this section views it as a sophisticated modification of the naive Bayesian approach to the reciprocal-reasoning regress. The iterative nature of that regress, that is, the "I think..., but then he thinks,..., so I should think...", etc. inner dialogue accompanying the reasoning, is adopted with one crucial change. In response to his "realizing" that his opponents have "discovered" his mixed-strategy prior, or some subsequent best reply, and have computed their best replies, a rational player does not assume for the purposes of calculation that they will be disposed to play those newly computed best replies. This is the naive assumption which leads to non-convergence of the regress. To be sure, the rational response is to assume that with their new "discovery" they, the opposition, will have individually altered in some way their assumptions of what you, one of their rational opponents, are disposed to play, and will have computed their best replies to this new estimate. This is what gives rise to the iterative character of the tracing procedure. The key to the LTP is that the single-iteration-change in any player's assumption regarding any of his opponents' strategic

dispositions is infinitesimal. The rational response to realizing an opponent has computed his best reply at some level of the regress is to make one's estimate of that opponent's strategic disposition a bit closer to that best reply, but only a bit. Exactly how this iterative modification of a rational estimate of an opponent's strategic disposition works will now be discussed.

The vertices of the polyhedron S, the collective strategy space, represent pure strategy n-tuples, $b_1, b_2, \ldots, b_k$. For almost all points $s \varepsilon S$ player i's best reply to $\bar{s}_i$ (his opponents' part of s) is a pure strategy since maximization of expected utility under conditions of certainty will almost always prescribe the playing of a unique pure strategy. Consequently, the collective best reply to almost all $s \varepsilon S$, is some vertex of the space S. The fundamental failure of the naive Bayesian approach to the reciprocal-reasoning regress is that the collective best reply to the prior probability distribution considered as a set of mixed strategies is not always an equilibrium point. Looking closer we can see why this is so.

The naive approach fails whenever the one-dimensional line segment joining the prior p to its collective best reply vertex b_v contains points to which b_v is not the collective best reply. Call this line segment α_1. The first point along α_1 where the collective best reply "switches"

from b_v to some other vertex b_v^1 may be called a <u>switching point</u>. Call this switching point w_1 and call the truncated version of α_1, $\bar{\alpha}_1$. The locus of all switching points for all $s \in S$ defines a set of switching "curves" in S. These curves partition S into sets of points, S_k, called "collective stability sets". These are such that the collective best reply to all members of a given set S_k is the pure strategy n-tuple represented by the vertex b_k. The switching curves themselves are in effect the intersections of two or more stability sets. (They are also subsets of S of measure zero.) This means that associated with almost all switching points are <u>two</u> collective best replies. In the example, if we assume that p is not a switching point, these two collective best replies may be considered as the "old" vertex b_v, associated with the points on $\bar{\alpha}_1$ (the "incoming" segment of α_1) and the "new" vertex b_v^1 associated with an "outgoing" line segment α_2.

Version II of the LTP may now be described. Given the objective prior probability distribution of each agent across his pure strategy options, an (almost always unique) path α_1 is defined joining p to its collective best-reply vertex b_v. The LTP reasoning iteration described above is such that each player iteratively reasons his way along α_1 until either the end point b_v is reached, in which case it is an equilibrium point and is the "solution" of the game, or until a switching curve is encountered. This would occur at

some switching point which we may label w_1. At w_1 another (almost always unique) line segment α_2 will be defined joining w_1 to its best reply vertex b_v^1. The LTP now prescribes that the players follow α_2 until either its end point b_v^1 is reached, or until another switching curve is encountered. If the latter obtains the preceding steps are to be repeated. Since this interpretation of the LTP is mathematically equivalent to Version I it will select a unique equilibrium point for almost all games and almost all initial prior probability distributions p. In short, for a given game Γ and a given prior p following the pathway $\alpha = \{\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \alpha_E\}$ almost always defines an equilibrium solution.

Four points need to be mentioned. First, it can be shown that in reasoning his way along α employing the iterative procedure described above, each player i computes his best reply in response to an "estimate" of the other players' strategic dispositions given by the formula:

$$(3.2) \quad \Pi_i = t\,\overline{s_i} + (1-t)\,\overline{p_i}$$

where $\overline{s_i}$ is the set of his opponents' best replies to their estimates Π_j, $j \neq i$. The parameter t is, of course, the same parameter Harsanyi employed in defining the family of games $\{\Gamma^t\}$ in Version I. Its function in this version of the LTP may be viewed as parametricizing the path α -- roughly speaking it is a non-linear measure of "distance along α

from the starting point". The pathway α is, in fact, the locus of points $\Pi(t)$ for $0 \leq t \leq 1$.

Secondly, the naive Bayesian approach discussed above is successful whenever $\alpha = \alpha_1$, i.e. whenever α is such that it crosses no switching curves for the game Γ. Each stable equilibrium point for Γ, then, can be seen to have associated with it a set of priors for which the naive approach succeeds.

Thirdly, given some prior probability distribution p and hence some unique pathway α'' (say), any switching point for α'' corresponds to a particular probability "estimate", $\Pi(t_q)$, against which some player i is indifferent between two or more best-reply strategies. Typically, however, all but one of these best replies ceases to be a best reply for player i to $\Pi_i(t)$ for $t > t_q$. For this reason and although the set of collective best replies to $\Pi(t_q)$ is not a singleton, we may safely treat α in the neighborhood of $\Pi(t_q)$ as simply a combination of an incoming segment for $t < t_q$, and an outgoing segment for $t > t_q$.

Finally, corresponding to those games and those priors for which the linear tracing procedure does not pick out a unique equilibrium point there are paths α which at some point $\Pi(t_q)$ branch into two or more outgoing segments for $t > t_q$, each branch corresponding to a "stable" collective best reply. We may call such points "singular" points. This happens when $\Pi(t_q)$ lies on two or more switching curves and so is a boundary point of three or more collective stability

If $\varepsilon = 0$ these games are the same as the games of the original family $\{\Gamma^t\}$. But for any $\varepsilon > 0$ these "starred" games are quite different from their counterparts, and have the very special property that for $t < 1$ each Γ^t_* has only one equilibrium point. Moreover, each player's strategy at that equilibrium point is a fully mixed strategy.

Without going into the finer points of Harsanyi's proofs, what he has shown is that if we let E^t_* be the set of equilibrium points of Γ^t_*, the set of points:

$$(3.5) \quad P^* = \{x : x=(t,s) \wedge \varepsilon[0,1] \wedge s\varepsilon E^t_*\}$$

is always a family of one-dimension, <u>fully</u> algebraic curves. From the definition of $\{\Gamma^t_*\}$ given above it is clear that as $\varepsilon \to 0$ the curves in P^* converge towards curves or parts of curves in P, the corresponding set in the LTP. More particularly, any path in P (for a given prior) which does not branch turns out to be the <u>limit curve</u> for some series of curves from P^* for different values of ε approaching zero (and given that same prior). The LoTP defines the equilibrium point of that limit curve as the "solution" to the game given that prior.

It is for those paths in P which <u>do</u> branch that the LoTP was designed. Because of the properties of the games $\{\Gamma^t_*\}$ and the derivative properties of the family of curves P^*, the limit curve corresponding to a path in P with two or more branches <u>always picks out only one branch</u>. The LoTP

how the team reaches its collective decision regarding the direction of optimal ascent) the team is almost always guaranteed to follow a unique path from its starting point (that is, the prior probability distribution) to a point in the terrain which has the character of a "local maximum". The local maximum is such that for all players it is the highest point in its neighborhood and so no further "direction of ascent" can be defined. When this happens the team has "solved" the mountain, and has reached a point corresponding to an equilibrium point of the game.[1]

3.3. The Logarithmic Tracing Procedure (LoTP)

3.3.1. Non-Technical Overview

We have seen that the linear tracing procedure is essentially a method for defining a family of paths in an abstract space of some sort based in some way on the strategy space and payoff vector for the game. Corresponding to almost all prior probability distributions is some unique member of this family of paths leading to a unique equilibrium point of the game called the "solution" of the game given that prior. Only when the specification of the prior does not pick out a unique path, or when the selected path "branches" at some point so more than one equilibrium point is reachable does the LTP not define a solution.

The Logarithmic Tracing Procedure (LoTP) is essentially a method for defining other closely related families of paths

in the same abstract spaces as the two versions of the LTP. The main feature of these new families is that no path has any branches. Consequently, a unique path and a unique equilibrium solution is always chosen by the specification of the prior probability distribution p.

Branches in the paths defined by either Version I or Version II of the LTP occur whenever some player has two or more "equally desirable" best replies, where "equally desirable" includes some consideration of the "stability" of the best reply. In Version I this stability requirement is satisfied whenever some small subset of the family of games $\{\Gamma^{t}\}$ defined in the neighborhood of some particular game Γ^{t_q} for $t>t_q$ all have two or more equilibrium points in common. In Version II this is the case whenever some player can respond to his "estimate" of what his opponents might play with two or more best replies each having the following additional property: It must continue to be a best reply <u>after</u> his opponents have taken whichever one he selects into account in <u>their</u> "estimates".

Points in the respective abstract space at which either ~~sort~~ of branching occurs may be called "singular points" for that version of the LTP. The LoTP is defined so that all paths avoid such singular points. This is accomplished by the straightforward strategy of ensuring that no player can ever have two best replies, let alone two best replies satisfying the "stability" requirement. This in

turn is accomplished by ensuring that only a fully mixed strategy can be a best reply. How this is done, and why it has the desired effect cannot be discussed without going into considerable technical detail, and so will be avoided in this introduction.

In general terms, then, the overall effect of the LoTP is the "smoothing" of the paths defined by either version of the LTP. The most important "smoothing" takes place in the neighborhood of those points where two or more collective best replies can be defined. This maneuver just so happens to avoid "singular points" and "branching", and so defines a unique equilibrium "solution" for all n-person non-cooperative games given some prior probability distribution.

As before, the non-technical reader may wish on a first reading to skip over the technical presentation which follows, and proceed to an examination of the illustrative example discussed below.

3.3.2. LoTP Version I

In his presentation of the LoTP Harsanyi introduces a second family of games $\{\Gamma_*^t\}$ with player i's payoff vector in game Γ_*^t defined by the equation:

$$(3.4)\quad v_i^*(s_i,\bar{s}_i;p,\varepsilon,t)=t\ u_i(s_i,\bar{s}_i)+(1-t)\left[u_i(s_i,\bar{p}_i)+\varepsilon\sum_{k=1}^{K_i}\log s_i^k\right]$$

If $\epsilon = 0$ these games are the same as the games of the original family $\{\Gamma^t\}$. But for any $\epsilon > 0$ these "starred" games are quite different from their counterparts, and have the very special property that for $t < 1$ each Γ^t_* has only one equilibrium point. Moreover, each player's strategy at that equilibrium point is a fully mixed strategy.

Without going into the finer points of Harsanyi's proofs, what he has shown is that if we let E^t_* be the set of equilibrium points of Γ^t_*, the set of points:

$$(3.5) \quad P^* = \{x : x=(t,s) \wedge \epsilon[0,1] \wedge s\epsilon E^t_*\}$$

is always a family of one-dimension, fully algebraic curves. From the definition of $\{\Gamma^t_*\}$ given above it is clear that as $\epsilon \to 0$ the curves in P^* converge towards curves or parts of curves in P, the corresponding set in the LTP. More particularly, any path in P (for a given prior) which does not branch turns out to be the limit curve for some series of curves from P^* for different values of ϵ approaching zero (and given that same prior). The LoTP defines the equilibrium point of that limit curve as the "solution" to the game given that prior.

It is for those paths in P which do branch that the LoTP was designed. Because of the properties of the games $\{\Gamma^t_*\}$ and the derivative properties of the family of curves P^*, the limit curve corresponding to a path in P with two or more branches always picks out only one branch. The LoTP

"legitimates" that branch, and selects its end point as the "solution" for the game given that prior.

In this version, then, the LoTP in effect defines P^* to be smoothed versions of the family of curves P. These smoothed curves have the key property that they do not branch. By carrying this property back into the set P by mathematical manipulation, unique equilibrium points can be selected for every prior and every game, branching or no.

3.3.3. LoTP Version II

This version of the LoTP treats it as a method for smoothing the family of pathways $\{\alpha\}$ in the neighborhood of the switching curves. To be absolutely precise, of course, the smoothing is carried out everywhere in S, but the only place where such smoothing is necessary is in the neighborhood of the switching curves where some player's best reply changes from one pure strategy to another as $\Pi(t)$ passes from one side of the switching curve to the other. We may call the family of "smoothed" pathways $\{\alpha^*\}$.

Corresponding to the family of smoothed pathways is a modified version of the old estimator function $\Pi(t)$. The exact form of this modified estimator function $\Pi^*(t,\varepsilon)$ is very complicated and need not concern us here. The two key effects of the modification, however, are plain enough. First, there are no longer any such things as switching curves. In place of each distinct curve where two best replies are defined for some player there are now "switching

neighborhoods" through which that player gradually replaces one best reply by the other. He does this by moving through intermediate best-reply strategies which are mixtures of the two as Π_i^* passes from one side of the old curve to the other. Secondly, the modification ensures that for no value of $t<1$ will any player consider a pure strategy to be a best reply. For estimates Π_i^* far away from any of player i's old switching curves, his computed best reply is arbitrarily close to the old pure-strategy best reply -- exactly how close being a function of the estimator parameter ε. However, for estimates Π_i^* <u>close</u> to an old switching curve, player i's best reply will be a mixed strategy intermediate between the two pure-strategy best replies corresponding to that old switching curve.

The overall effect of these changes is that in the neighborhood of the old switching curves the pathways $\{\alpha\}$ are smoothed so as to have a continuous derivative. More importantly, since the smoothed pathways have no sudden changes in direction no branching of the sort which plagued the LTP can occur.

Similarly to what the LoTP Version I did for branching curves in P, this interpretation of the LoTP takes any branching pathway from $\{\alpha\}$ and "legitimates" that branch which corresponds to part of the limit pathway for $\varepsilon \to 0$ of the corresponding member of $\{\alpha^*\}$. Unique legitimated pathways are always thereby defined for all priors and all games,

and the endpoints of those pathways are the same equilibrium points selected by Version I as the "solution" to the game given that particular prior.

3.4. An Example

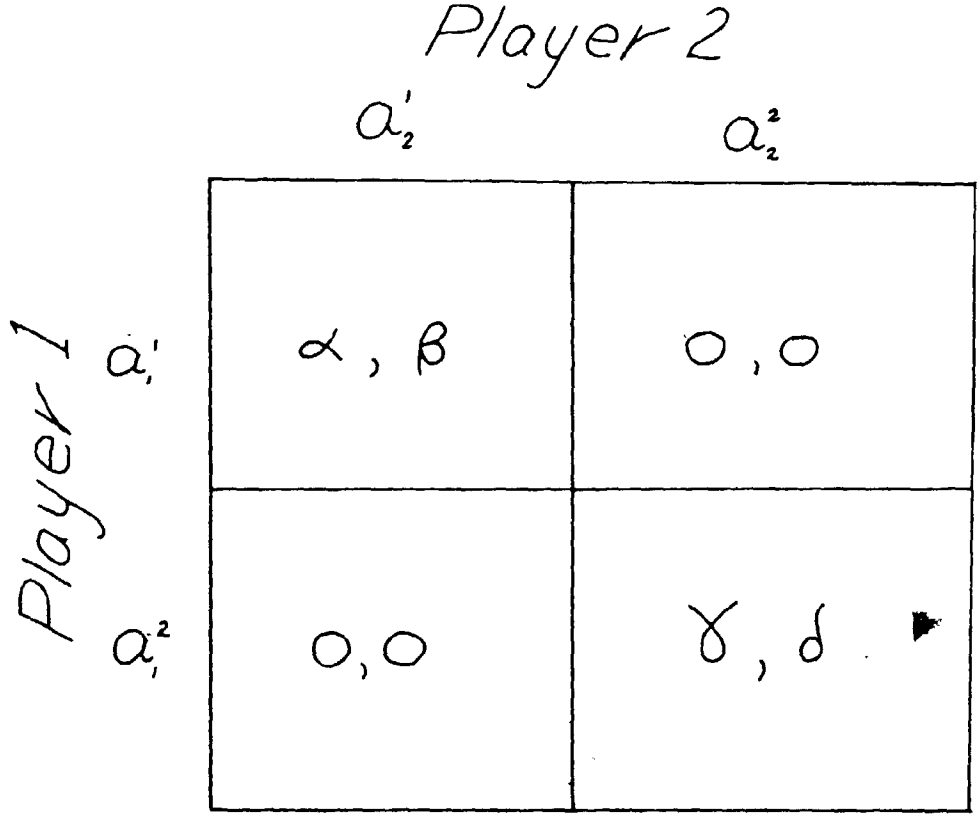
3.4.1. Non-Technical Overview.

One of the most simple and yet most important forms of two-person non-cooperative games studied in the game theory literature is that whose normal form is representable by the game matrix of Figure 3.1. In any game of this form there are two players each having two strategy options. The context of interaction is defined to be such that given a choice by either player the rational choice by the other is unique and is in equilibrium with the given strategy of his opponent. In the absence of such information neither player has an obviously preferred strategy.

This class of games present non-cooperative theory with a particularly recalcitrant problem. What is the rational course of action when the players must choose simultaneously and yet are not in a position to co-ordinate their choices so as to avoid the mutually disastrous off-diagnoal payoffs?

Early attempts to define solutions for games of this sort tried to justify a particular choice by the individual players by arguing for the superiority of one of the two equilibrium outcomes. It is clear, for example, that if $\alpha > \gamma$ and $\beta > \delta$ both players prefer the outcome (α, β) to the

Figure 3.1. A Set of Games


Player 2
a_2^1
a_2^2
Player 1
a_1^1
a_1^2
α, β
0, 0
0, 0
γ, δ

outcome (γ,δ). It would seem rational, then, to assume that (α,β) is in some sense the "target" for both players. Under this assumption about the other player the choice of the equilibrium strategies a_1^1 and a_2^1 could easily be justified, and the "solution" to the game thereby defined.

Whatever the merits of this approach, there are classes of games with values for α,β,γ and δ which block the analysis. One of the most famous of these is the so-called "Battle of the Sexes" game in which neither of the diagonal equilibrium points is jointly preferred. The standard interpretation is something like the following. The players are husband and wife, and the context is the mutual selection of the evening's entertainment out of two competitors -- a boxing match and the ballet. By the rules of the game or by some contrived set of circumstances, the individual selections must be made simultaneously with no communication between the players, and the result must be a unanimous decision or else the couple will be forced to spend the evening at home, an eventuality neither prefers to either of the two alternatives. However, as it happens the wife prefers the first alternative, the boxing match, to the second, the ballet, and the husband prefers the second to the first. What is each to do?

In order to illustrate some of the technical issues raised during the discussion of the tracing procedure I want to explore this example in a bit more depth. The reader

might find it helpful to keep the Battle of the Sexes interpretation in mind while working through the analysis. I will return to it when the technical aspects of the example have been elucidated.

3.4.2. The Tracing Procedure Applied to the Example

The collective strategy space for any member of the set of games of Figure 3.1 is the unit square, since the strategy space for each player is simply the unit line representing the probability mix between the first and second strategy options.[2] Figure 3.2 shows such a strategy space.

Now any game of the form of Figure 3.1 has three equilibrium points, e_1, e_2, e_3, defined by the following:

$$(3.6)\quad \begin{aligned} e_1 &= (a_1^1, a_2^1) \\ e_2 &= (a_1^2, a_2^2) \\ e_3 &= (s_1^*, s_2^*) \end{aligned}$$

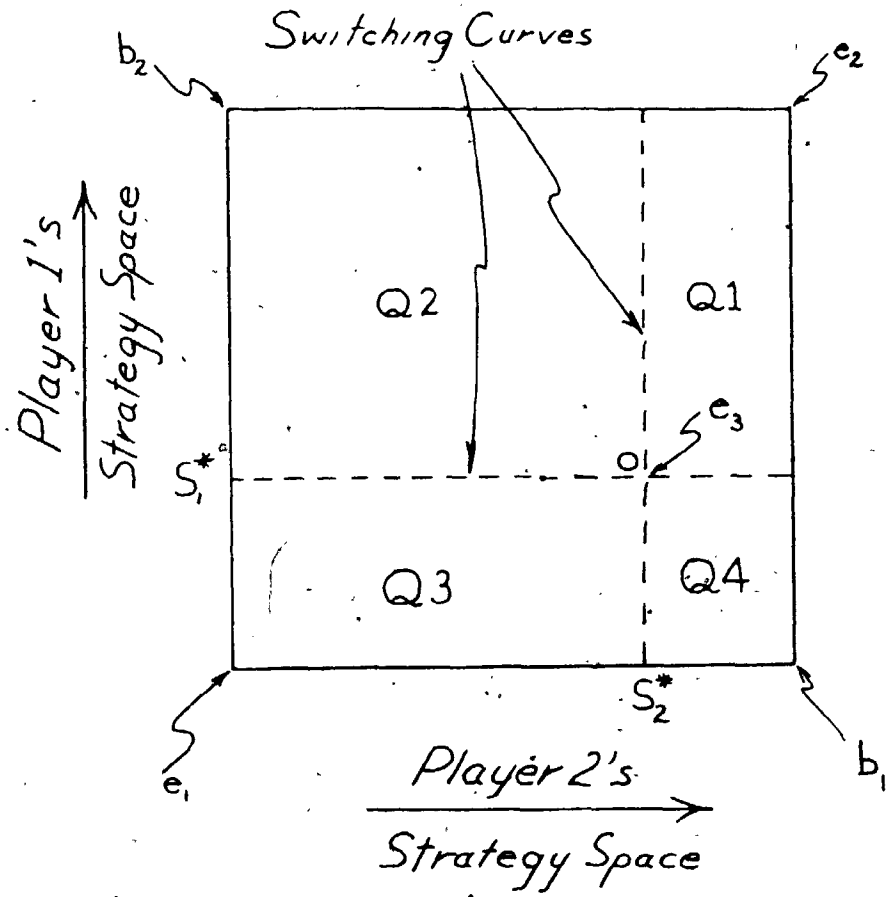
where, $$(3.7)\quad s_1^* = (s_1^{1*}, s_1^{2*}) = \left(1 - \frac{\beta}{\beta+\delta}, \frac{\beta}{\beta+\delta}\right)$$

$$s_2^* = (s_2^{1*}, s_2^{2*}) = \left(1 - \frac{\alpha}{\alpha+\gamma}, \frac{\alpha}{\alpha+\gamma}\right)$$

These are noted in Figure 3.2.

In Figure 3.2 the dotted lines parallel to the axes are the switching curves. Player 1's is parallel to his axis, player 2's to his axis, since both switching curves are

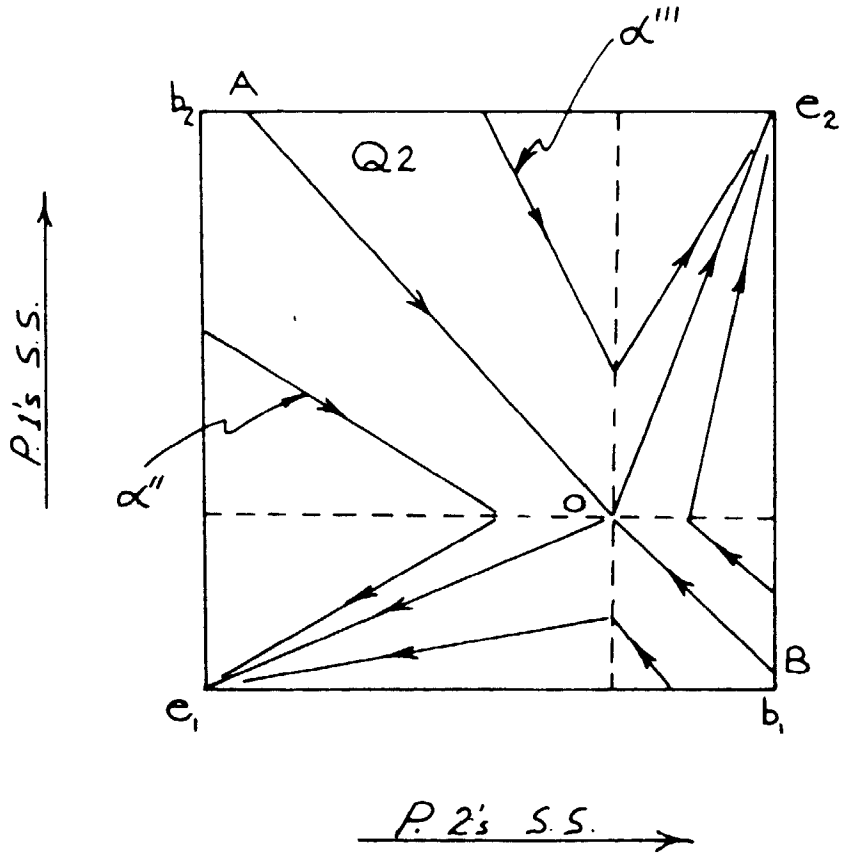
Figure 3.2. The Collective Strategy Space


Switching Curves
b_2
e_2
Player 1's Strategy Space
Q2
Q1
e_3
S_1^*
Q3
Q4
S_2^*
e_1
b_1
Player 2's Strategy Space

To make matters even worse there is a third viable reasoning option according to the LTP. The collective strategy e_3 is not itself a "conventionally-stable" equilibrium point since a very slight change in either player's strategy mix away from e_3 causes the other to choose some pure strategy as his best reply. Nevertheless, the way the LTP has been defined, in response to a prior strategy mix represented by e_3 each player i might rationally choose his own contribution to that strategy, $(e_3)_i$, as his best reply to his opponent j's contribution, $(e_3)_j$. Since this response to the estimator output $\Pi=e_3$ is not denied by the LTP, the output $\Pi(t)$ might legitimately remain equal to e_3 until $t=1$. Consequently, e_3 itself must be considered as a third viable "solution" option for priors lying on A0B.

The LoTP, of course, is designed to choose from among these three options whenever point 0 is encountered in the players' reasoning. As discussed above the technical details of that procedure are such that each member of the family of pathways $\{\alpha\}$ is "smoothed" in the neighborhood of the switching curves. Figure 3.4 represents a few of the members of the smoothed family $\{\alpha^*\}$ with a much exaggerated smoothing parameter ε. The most important thing to note in this figure is the behavior of the pathway corresponding to priors lying on the previously troublesome A0B. For example, consider a collective strategy prior far away (relative to ε) from the point 0 but lying on A0. The logarithmic

Figure 3.3. A Few Members of $\{\propto\}$


α'''
A
b2
e2
Q2
P.1's S.S.
α''
O
B
e1
b1
P.2's S.S.

in Figure 3.3 to illustrate this. (Note the switching points at which the pathways instantaneously change direction.)

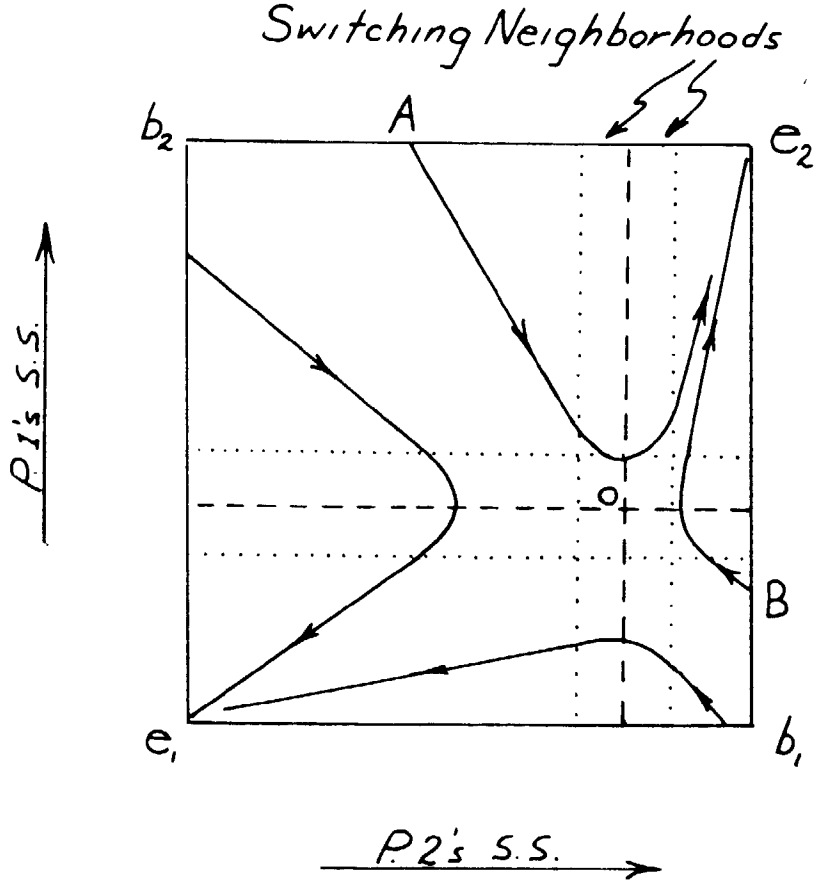
The second thing to note is that the specification of a prior in Q2 can have one of two possible outcomes. If the prior lies to the upper right of the line segment A0 (the extension of $b_1 0$) the result is the selection of e_2 as "solution". Given a prior to the lower left of that line the "solution" is e_1. Should a prior lie <u>on</u> A0, branching in the resultant "reasoning" occurs at e_3, and both equilibrium points are reachable.[3]

Starting at a prior on A0 the players begin reasoning in the usual fashion by selecting the vertex b_1 as their collective best reply. However, at that value for t, say t_0, when the collective estimator function $\Pi(t_0)$ determines the collective strategy disposition to be the mixed equilibrium strategy e_3, each player discovers that <u>any</u> response fulfills the requirement of a best reply. More importantly, two of these responses <u>also</u> satisfy the "stability" requirement for a viable reasoning pathway: The players might jointly choose e_2 as the collective best reply and thereafter reason along $0e_2$, or they might jointly choose e_1 and thereafter reason along $0e_1$. Since both of these "stable" collective best replies are available to the players at point 0, the LTP fails to define a solution whenever the prior is such that the selected pathway passes through that point. This is the case whenever the prior lies on A0B.

To make matters even worse there is a third viable reasoning option according to the LTP. The collective strategy e_3 is not itself a "conventionally-stable" equilibrium point since a very slight change in either player's strategy mix away from e_3 causes the other to choose some pure strategy as his best reply. Nevertheless, the way the LTP has been defined, in response to a prior strategy mix represented by e_3 each player i might rationally choose his own contribution to that strategy, $(e_3)_i$, as his best reply to his opponent j's contribution, $(e_3)_j$. Since this response to the estimator output $\Pi=e_3$ is not denied by the LTP, the output $\Pi(t)$ might legitimately remain equal to e_3 until $t=1$. Consequently, e_3 itself must be considered as a third viable "solution" option for priors lying on AOB.

The LoTP, of course, is designed to choose from among these three options whenever point 0 is encountered in the players' reasoning. As discussed above the technical details of that procedure are such that each member of the family of pathways $\{\alpha\}$ is "smoothed" in the neighborhood of the switching curves. Figure 3.4 represents a few of the members of the smoothed family $\{\alpha^*\}$ with a much exaggerated smoothing parameter ε. The most important thing to note in this figure is the behavior of the pathway corresponding to priors lying on the previously troublesome AOB. For example, consider a collective strategy prior far away (relative to ε) from the point 0 but lying on AO. The logarithmic

Figure 3.4. A Few Members of $\{\alpha^*\}$


Switching Neighborhoods
A
b_2
e_2
B
e_1
b_1
o
P1's S.S.
P2's S.S.

estimator will in effect have the players begin by reasoning towards b_1 in the same fashion as before. In general, however, as the estimator output approaches the mixed-strategy equilibrium e_3 one of the "switching neighborhoods" will be encountered before the other. In Figure 3.4 player 1's switching neighborhood is encountered first. Therefore, as the estimator output approaches e_3 and thus what used to be the intersection of the players' switching curves, player 1 begins to replace his "old" best reply with the new one. Player 2, of course, has not "begun" his switch and in this case never will. As a result, the direction of the collective best reply "bends" away from b_1 and towards e_2.

Although Figure 3.4 shows this "bending" and the early encounter of player 1's neighborhood for an exaggerated value of ε, both of these things obtain for this game for any value of $\varepsilon > 0$.

Since in the limit as $\varepsilon \to 0$ the smoothed pathway Ae_2 converges to the piece-wise algebraic path $A0\text{-}0e_2$, e_2 is selected as the "legitimate" solution for any prior lying on A0. Of course, the same holds for priors lying on B0.

In general, e_2 is selected for priors lying on A0B whenever the game is such that $\gamma\delta > \alpha\beta$, while e_1 is selected when $\alpha\beta > \gamma\delta$. Should $\alpha\beta = \gamma\delta$ the LoTP selects the unstable equilibrium point e_3 as the solution of the game given such a prior.[4]

3.4.3. Interpretation and Analysis

Returning now to the Battle of the Sexes game described above, these results can be illustrated. The point of the tracing procedure is to pick out equilibrium strategy choices for both the husband and wife: Both must choose to go either to the ballet or to the boxing match. It does this on the basis of two types of information: an "objective" prior disposition of each player towards the two alternatives, and the relative importance of the two alternatives to each player. Consider a "battle" where the utility assignments are as follows (see Figure 3.5):

$$\begin{aligned} \alpha &= 1 \\ \beta &= 6 \\ \gamma &= 3 \\ \delta &= 4 \end{aligned} \qquad (3.8)$$

The strategy space and family of smoothed pathways $\{\alpha^*\}$ for this game are as shown in Figure 3.6. The results of applying the LoTP are as follows:

1) If the prior lies on or to the upper right of AOB, both parties should choose to go to the ballet.

2) If the prior lies to the lower left of AOB, both parties should choose to go to the boxing match.

Of course, until the prior probability distribution is specified the "solution" to this game has not really been defined. Nevertheless, what it means to be a "solution" is clear, and we can be confident that no matter what prior is

Figure 3.5. A "Battle of the Sexes" Game

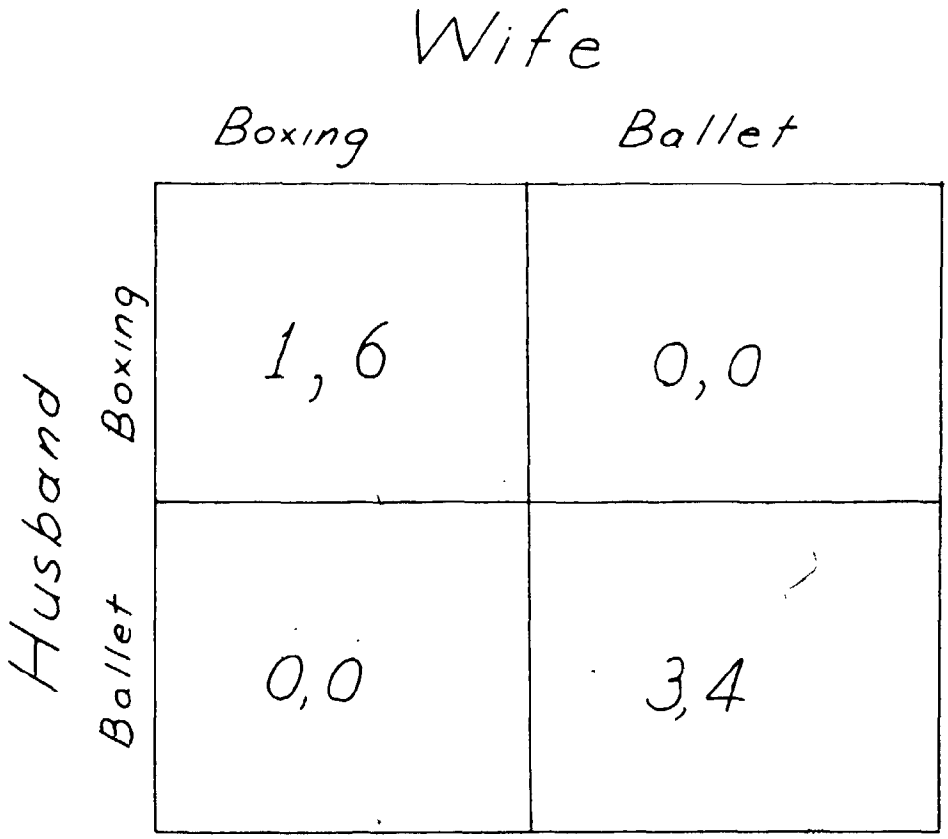

Wife
Boxing
Ballet
Husband
Boxing
Ballet
1, 6
0, 0
0, 0
3, 4

Figure 3.6. The Strategy Space for the "Battle of the Sexes"

n-person game in normal form and a unique point in the collective strategy space for that game (i.e. the "prior") the algorithm will always be able to arrive at a unique equilibrium point of the game (i.e. the "solution").

2) Normative Interpretability: Given the usual normative interpretation of the normal form for n-person non-cooperative games, the input data to the algorithm and the data generated by the algorithm are consistently interpretable as what the rational agents playing the game want, know and decide. More particularly, each agent is a utility maximizer, and given the iterative nature of the algorithm, each step involves the calculation of best replies.[2]

3) Psychological Interpretability: The "reasoning processes" reconstructed by the algorithm are those of actual human beings, although not all people need reason in exactly this way and some parts of the algorithm may be idealizations of actual psychological processes.[3]

It is clear, then, that four categories of criticisms could be levelled at the tracing procedure, three to the effect that it fails to meet one of the criteria of adequacy, and one more to the effect that these are the wrong criteria of adequacy for a proposed theory of rational interaction.

Harsanyi's skill as a mathematician may be taken for granted. Therefore, I will not question here whether he has satisfied the formal criterion. However, I do want to argue that he has not satisfied the other two criteria, Normative Interpretability and Psychological Interpretability. This will be done in the next two sections. In Chapter VIII I will examine whether or not these three criteria are what a theory of rational interaction must meet.[4]

Of course, it is Harsanyi's concern with the positive interpretation of his model which makes his approach even

specified, a unique prescription can be made.

Consider a prior representing the players' uncertainty by an equipartition assumption. If a rational representation of each player's ignorance of the other's predispositions is such an assignment of a (.5,.5) prior, the LoTP prescribes for this game the selection of the ballet. On the other hand, if a "psychological predisposition" is what determines the "objective" prior, and if each party is known to be psychologically predisposed towards altruism, the objective prior would be the vertex b_1. Taking this as a prior the LoTP again prescribes the rational non-cooperative choice by each party to be the ballet.

FOOTNOTES

1. This "hill-climbing" model of the tracing procedure is the basis for a generalization allowing for other ways of combining the independent "suggestions". (See below Chapter IV, section 4.2.3.5.) An interesting study of the control-theoretic approach to games is found in Blaquiere et al., Quantitative and Qualitative Games (1969).

2. In general, the strategy space for player i will have (K_i-1) dimensions, since any mixed strategy can be specified by specifying probabilities (≤ 1) for all but one of the pure strategy options. The remaining probability if fixed by the stipulation that the sum of the probabilities equals 1.

3. The same comments apply to Q4 and priors lying on the line B0, the extension of $b_2 0$.

4. This result is not that intuitive and Harsanyi owes us an argument to show that this is not simply spurious. As the interpretation in terms of "switching neighborhoods" shows, when $\alpha\varepsilon = \gamma\delta$ and the prior lies on A0B, both players' switching neighborhoods are "encountered" at the same value for the parameter t, regardless of what value is assigned to ε. The strange result of this "simultaneous encounter" is the "slowing down" of the move towards e_3. This is because as $\Pi^*(t)$ approaches e_3, the best reply to $\Pi^*(t)$ also approaches e_3, and so the "estimate" is modified only slightly. When t becomes equal to 1, $\Pi^*(t)$ is equal to e_3. Therefore, the limit curve of this sort of α^* picks out e_3 as its end point and the "solution" given that prior.

CHAPTER IV

CRITIQUE OF THE TRACING PROCEDURE

4.1. Introduction

It was suggested in Chapter II that Harsanyi's approach to normative issues generally and to game theory in particular has always been a unique combination of formalism and a desire for psychological meaningfulness. This is what makes his approach to rationality a "dual-interpretation" approach.

In game theory he has adopted the normatively-motivated but purely formal constraint of the normal form of representation, while at the same time insisting that any formal algorithm proposed as a solution to the normative problem of interdependent choice also admit of a psychological or "positive" interpretation. That is, he requires that the proposed algorithm model the reasoning processes of actual intelligent human beings, and that the model function in an explanatory context.

In short, Harsanyi's approach specifies the following three criteria of adequacy for normative solution concepts in game theory:[1]

1) Formal Existence and Uniqueness: Given input data consisting of the game vector for any

n-person game in normal form and a unique point in the collective strategy space for that game (i.e. the "prior") the algorithm will always be able to arrive at a unique equilibrium point of the game (i.e. the "solution").

2) Normative Interpretability: Given the usual normative interpretation of the normal form for n-person non-cooperative games, the input data to the algorithm and the data generated by the algorithm are consistently interpretable as what the rational agents playing the game want, know and decide. More particularly, each agent is a utility maximizer, and given the iterative nature of the algorithm, each step involves the calculation of best replies.[2]

3) Psychological Interpretability: The "reasoning processes" reconstructed by the algorithm are those of actual human beings, although not all people need reason in exactly this way and some parts of the algorithm may be idealizations of actual psychological processes.[3]

It is clear, then, that four categories of criticisms could be levelled at the tracing procedure, three to the effect that it fails to meet one of the criteria of adequacy, and one more to the effect that these are the wrong criteria of adequacy for a proposed theory of rational interaction.

Harsanyi's skill as a mathematician may be taken for granted. Therefore, I will not question here whether he has satisfied the formal criterion. However, I do want to argue that he has not satisfied the other two criteria, Normative Interpretability and Psychological Interpretability. This will be done in the next two sections. In Chapter VIII I will examine whether or not these three criteria are what a theory of rational interaction must meet.[4]

Of course, it is Harsanyi's concern with the positive interpretation of his model which makes his approach even

be considered as a "rationally playable" mixed strategy.[9] On the one hand, if the prior would never be a rational strategy to play, but only to "assume", the rational player has the same set of contradictory beliefs here that he has at any later point in the regress -- he is asked to believe that in some sense his opponents are "likely" to play the prior whereas he also knows they will not. I take it this is still unacceptable.

The other option, therefore, seems more likely. Unlike estimates $\Pi(t)$ for values of $t>0$ the collective prior is in fact a "rationally playable" collective mixed strategy. In fact, it is the best one available to the players were the reasoning regress to be halted before it got started. But if this is the case, that is, if the notion of an objective prior strategy entails that it is "rationally playable", then we must surely ask in what sense of "rational" this obtains.

On the one hand, we might reasonably expect that that strategy, the prior, like any other a rational Bayesian would think of playing, is itself a best reply to some sort of assumed probability distribution, say p', across his opponents' strategy options. However, this cannot be the case for it entails one of the following:

Either: 1) That each opponent has chosen to play according to p' in some non-Bayesian fashion, which is denied by mutual rationality,

or: 2) That the prior mix p', is itself a collective best-reply response to an assumed probability

Whatever problems this approach has, at least it has the following very important characteristic: at no point in the regress (once it starts) does any player ever assume his opponent is disposed to play anything other than a best reply to what he believes his opponents are likely to play. Of course, what data the opponent takes into account in computing his best reply varies from stage to stage in the iteration, but that does not bear on the main point. This point is simply that the estimates each player arrives at regarding his opponents' strategic dispositions are always consistent with his knowledge that each player is rational and hence will only play some best-reply strategy no matter what the particulars are of the assumptions he makes at any given time about his opponents' dispositions.

Unfortunately, the tracing procedure does not have this property and the result is disastrous. On the one hand, at any point (i.e. value for the parameter 't') in the regress, each player i computes his own best reply in response to an "estimate" of his opponents' strategic disposition given by $\Pi_i(t)$. For the purposes of computation, then, it is rational for player i to assume that his opponents "might" play according to that estimate. At the very least it must be conceivable that given their present beliefs his opponents might play according to that estimate. Otherwise, there would be no point at all in computing a best reply to it.

On the other hand, it is also inconceivable that a

rational-Bayesian will play anything other than a best reply once he has any estimate at all of his opponents' likely collective strategy. This is reflected in Harsanyi's own admission that if for some reason the reasoning regress is interrupted at some value for t (say at t_I), the rational thing for each player i to do would be to play his best reply to the most recently calculated estimate $\Pi_i(t_I)$.

Now, since by the assumption of mutual rationality each player knows his opponents are rational, he also knows that if forced to choose, each will only play one of his best-reply strategies. But in general, the estimate the tracing procedure provides each player with is not the collection of his opponents' best replies given the beliefs he knows them to be holding. Rather, it is some strategy part way between that collective best reply and the priors. The rational player knows his opponents' beliefs, of course, and therefore knows that the estimate is not their collective best reply.[6]

In short, on the one hand Harsanyi contends that it is "rational" in some sense for a player to believe for the purposes of calculation that his opponents will play according to the estimator $\Pi(t)$. On the other hand, by the assumption of mutual rationality all players know that they will only play their best replies to $\Pi(t)$ no matter what the circumstances. Since for $t<1$, $\Pi(t) \neq$ a collective best reply to $\Pi(t)$, Harsanyi would have his rational player believe at the same time and in connection with the same

problem both a proposition "p" and its contradictory, "not-p". This is unacceptable.[7]

4.2.3. On the Notion of an "Objective" Prior

4.2.3.1. Introduction

As many people have pointed out, Harsanyi among them, Bayesians fall into two camps -- those who interpret prior probabilities as purely "subjective", and those who claim that an "objective" specification of prior probabilities is possible.[8] Whatever the larger ramifications of this dispute, it is clear that Harsanyi's tracing procedure requires that priors be "objective" in the sense that all rational players can compute them from the game vector alone. However, what else is required of the "prior" is not at all clear.

To help clarify this notion, I want to explore two possible answers to two questions: "Is the prior a 'rationally playable' collective mixed strategy?" and, "If so, in what sense of 'rational' is it 'rationally playable' Bayesian or non-Bayesian?" I argue that none of the three possible responses to this pairing of questions could be satisfactory for Harsanyi, and so conclude that his tracing procedure cannot get off the ground.

4.2.3.2. The dilemmas

The first question which must be asked is whether the prior probability distribution for a given player can

be considered as a "rationally playable" mixed strategy.[9] On the one hand, if the prior would never be a rational strategy to play, but only to "assume", the rational player has the same set of contradictory beliefs here that he has at any later point in the regress -- he is asked to believe that in some sense his opponents are "likely" to play the prior whereas he also knows they will not. I take it this is still unacceptable.

The other option, therefore, seems more likely. Unlike estimates $\Pi(t)$ for values of $t>0$ the collective prior is in fact a "rationally playable" collective mixed strategy. In fact, it is the best one available to the players were the reasoning regress to be halted before it got started. But if this is the case, that is, if the notion of an objective prior strategy entails that it is "rationally playable", then we must surely ask in what sense of "rational" this obtains.

On the one hand, we might reasonably expect that that strategy, the prior, like any other a rational Bayesian would think of playing, is itself a best reply to some sort of assumed probability distribution, say p', across his opponents' strategy options. However, this cannot be the case for it entails one of the following:

Either: 1) That each opponent has chosen to play according to p' in some non-Bayesian fashion, which is denied by mutual rationality,

or: 2) That the prior mix p', is itself a collective best-reply response to an assumed probability

> distribution across each player's opponents' strategies. However, this in turn entails propositions analagous to either 1) or 2), and so either denies the assumption of mutual rationality or propagates an infinite backwards regress of the naive Bayesian sort.

It would seem plausible, therefore, that if the prior is a "rationally playable" strategy, it is so in some not-straightforwardly-Bayesian sense of "rational strategy". At first glance this appears as a classic case of special pleading, especially for a Bayesian of Harsanyi's conviction. Could it be that Harsanyi is showing us that the foundation of a Bayesian theory of non-cooperative games is, (horror of horrors) a minimax decision rule, or (worse yet), a Laplacean equipartition assumption? Surprising as it may seem, this would appear to be the case. The next section is devoted to an examination of Harsanyi's likely response to this line of analysis.

4.2.3.3. Harsanyi's notion of "uncertainty"

In interpreting the values computed by the strategic estimator $\Pi(t)$ Harsanyi introduces a new notion of "uncertainty". To distinguish it from the more familiar notion let us call it "uncertainty$_2$". This notion is the key to Harsanyi's most likely response to the criticisms against the tracing procedure which have been raised so far in this chapter.[10]

We are, of course, accustomed to interpreting any probability distribution across states of nature or an opponent's pure strategy options as reflecting our own

"uncertainty" about which state of nature or which state of mind actually obtains. For example, in two-person zero-sum games the theory is designed so as to deny rational knowledge of the intended pure strategy. This is done through the artificial device of producing only a randomized mixed strategy as "final" result. The rest is left up to "fortune", and the epistemic constraints on the normal form say nothing about the players having equal knowledge of the results of the coin toss, or whatever. This form of "uncertainty", strategy ignorance consistent with the assumptions of the normal form, may be called "uncertainty$_1$". It is what the Bayesian approach to decision theory is all about.

Uncertainty$_2$, however, is quite different. In Harsanyi's words, uncertainty$_2$ arises in trying:

> . . . to assess the probability p_i^k that any strategy a_i^k should be player i's best reply to the strategies he might expect the other players to use. This means that, other things being equal, the prior probability p_i^k assigned to any pure strategy a_i^k will be greater the greater the range of possible situations in which a_i^k will be a best reply in the game.[11]

What Harsanyi seems to be saying is that we know that all rational players will play only best replies, since they are rational Bayesians, but at the beginning we do not know which among their best replies they will in fact play, since at the beginning of the regress we do not know what strategy assumption regarding their opponents' choices they are responding to. We do not know this because we do not

yet know how they "assess the situation". Nevertheless, we do know the way in which they go about reaching an assessment; they simply estimate their opponents' likely collective strategy! But since at the beginning of the regress no one has any idea at all about the details of his opponent's assessment, everyone should rationally assume total ignorance about the particulars of each player's belief. That is, in computing the prior each player can rationally assume that each opponent is selecting a best reply to some assumed collective strategy, but which one cannot be known. This ignorance about an opponent's "assessment of the situation" is reflected in uncertainty$_2$. Ignorance about his actual strategy is reflected in uncertainty$_1$, whatever the grounds for that ignorance (eg. he might be randomizing).

From the above quotation and this analysis it would appear that Harsanyi is committed to the resolution of uncertainty$_2$ by an equipartition assumption across each player's range of possible assessments. This is almost inescapable, since only then does each player's prior reflect in the most direct fashion "the range of possible situations in which [each pure strategy is] ... a best reply in the game". The rational player assumes that every one of each opponent's possible assessments is equally likely and therefore that the prior probability of that opponent playing any particular pure strategy best reply is proportional to the size of that strategy's stability set.

This shift from the level of an opponent's strategies to the level of his "estimates" or assessments of his opponents' collective strategy, introduces one more level of complexity into the Bayesian framework. The subtleties of the move are worthy of Harsanyi's sophistication. Nevertheless, I think the move fails to solve the problems besetting the notion of an "objective prior". The way I see it there are three interconnected reasons for this failure.

First, once an algorithm is provided to each player for computing the priors, all players do know how their opponents are "assessing the situation". There is no value for $0 \leq t \leq 1$ for which any player is in any sense "ignorant" of how his opponents are reasoning -- they all assess things in exactly the same way. This is where game theory has always floundered. The assumption of the transparency of reasoning gives each player a prohibitive knowledge of his opponent's reasoning.

Secondly, to avoid this problem Harsanyi has shifted from one level of epistemic ignorance -- the level of the priors -- to the next highest level in an infinite hierarchy -- the level where the algorithm which computes the prior gets its input, i.e. to the level of assessments. The priors are then computed automatically from some assumed distribution across each player's possible assessments. The retreat is potentially infinite, of course, but Harsanyi has arbitrarily stopped at the level of assessments.

We must press the issue, though, and ask what reason there is to think there is not some "objective" algorithm which computes the "rational" assessment given some other sort of probabilistic input. There is no reason I can see to make this assumption, and so Harsanyi's choice of a stopping point for the regress would seem to be entirely arbitrary.[12]

The third objection which must be raised concerns how Harsanyi has chosen to provide his prior-computing-algorithm with its input, the pre-reasoning assessments, given that there is no algorithm at the next highest level. His straightforward proposal is to assume that each possible assessment is equiprobable! That is, he employs an equi-partition assumption at the level of pre-reasoning assessments. Combined with the two previous arguments against adopting this approach to defining an objective prior, the need to resort to this thoroughly ad hoc proposal is sufficient to throw into doubt the whole notion of an "objective prior".

4.2.3.4. Summary of argument so far

The arguments so far have attacked Harsanyi's notions of a "rational" estimate and a "rational" prior on the grounds that if a "rational agent" is interpreted as being a "best-reply-playing-Bayesian", then the notions of "prior strategy" and "estimated strategy" involve him in holding contradictory beliefs. On the other hand, any extension to that interpretation is ad hoc. Harsanyi's attempt to split the horns of this dilemma by introducing

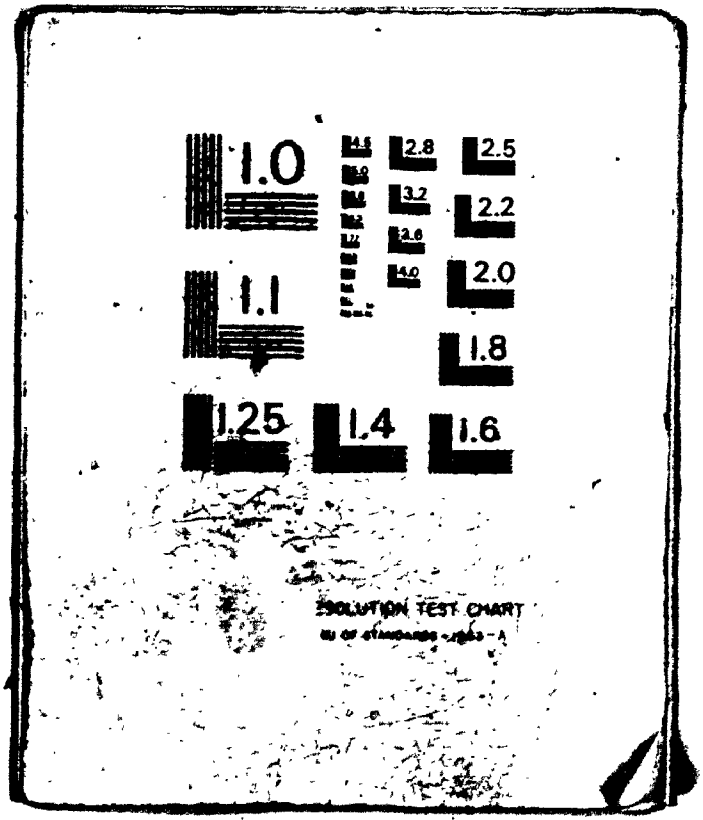
the notion of uncertainty$_2$ applicable to each player's ignorance of his opponents' "assessment of the situation" fails (a) because it is founded on a blatantly ad hoc equipartition assumption, and (b) because it does not reflect the belief state of the rational-Bayesian in possession of the tracing procedure.

The following section offers a more general criticism of the tracing procedure on the grounds that the estimator it provides to the rational player is itself ad hoc.

4.2.3.5. The tracing procedure generalized

If for the sake of discussion we loosen up the interpretability requirements on intermediate calculations in an iterative approach to games with assumed "priors", the linear tracing procedure can be seen more easily as simply one member of a family of functions $\Phi = \{\phi\}$, each ϕ mapping the collective strategy space (S) into the real line (R) in such a way that all and only equilibrium points of the game are local maxima in $\phi(S)$. The prior probability distribution for the game may then be regarded as a "test particle" placed in the "potential field" defined by ϕ. The set of all such "test particles" trace out the lines of flux for the field, and it can be seen from the above definition of Φ that these lines all converge on equilibrium points. Alternatively, each ϕ may be regarded as the measure of altitude employed by the mountain climbing team discussed in section 3.2.4.[13]

2

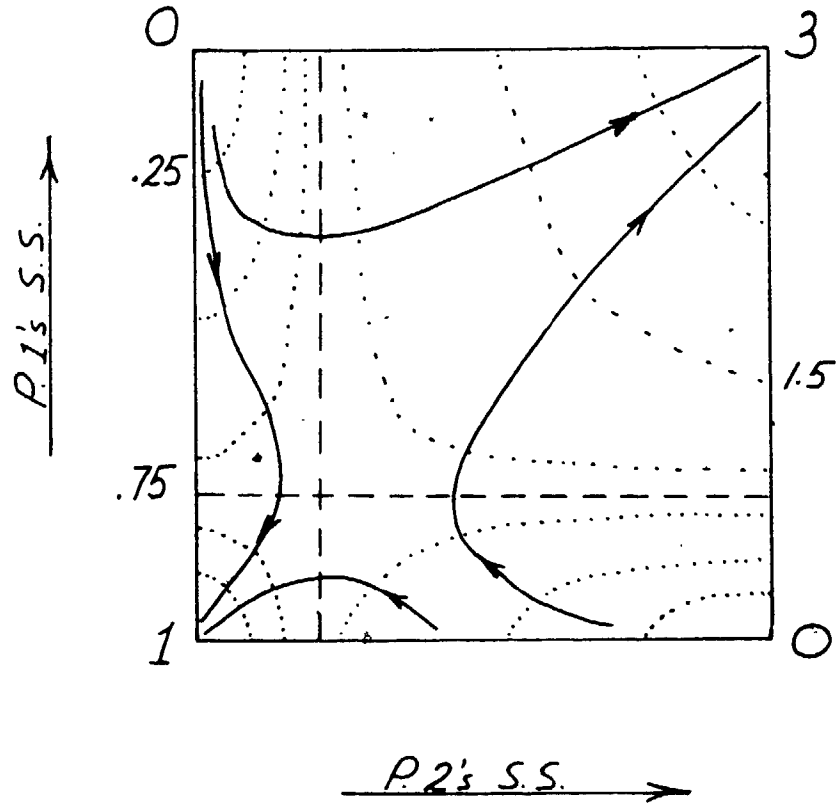


Now, to be sure, some interpretation must be given to any of these functions being proposed as the definition of the "field" through which rational players reason their way towards an equilibrium point. Clearly, for example, the potential should in some way be a function of both the utility functions of the players and the game vector. What more should be required of such functions so that the lines of flux represent idealized practical reasoning processes is, I take it, the problem of non-cooperative games when approached in a Bayesian fashion.

To illustrate the wide range of such ϕ open to consideration, consider again the simple 2 x 2 "Battle of the Sexes" game given in Figure 3.5. As discussed earlier, the strategy space for this game is represented by the unit square. Figure 3.6 showed this strategy space with "lines of flux" representing the tracing procedure's definition of a potential field ϕ^t. Figure 4.1 is another such diagram of the strategy space, but with a different potential function represented. Its equipotential contours and lines of flux correspond to a function ϕ^1 representing player 1's expected utility. That is, corresponding to each point $s \in S$ is the utility player 1 would receive should both players play their part of the collective mixed strategy s. A plausible normative, yet non-normal form, interpretation of this "field" is that each line of flux is the locus of increasingly superior collective strategies (from player 1's

Figure 4.1. Potential Field Representing Player 1's Utility Function, $u_1(s)$

C



perspective) given that:

1) a starting point has been defined

2) player 1 is in control of the collective strategy

3) player 1 is ignorant of the values for ϕ in all but the immediate neighborhood of the strategy point(s) under consideration at any particular time. That is, he knows the expected utility of s, and the gradient of the field at s, but is ignorant of all else including the overall structure of the game vector.[14]

Figure 4.2 shows a similar potential field representing player 2's expected utility. It can be interpreted in a similar fashion.

Figure 4.3 shows the field representing the product of the expected utilities for every collective strategy. In this case, each line of flux may reasonably be interpreted as representing the locus of optimum strategy improvements in a cooperative game where the players are concerned with maximizing the product of their expected utilities. Of course, they are under the same conditions of ignorance spelled out above.

For non-cooperative games, of course, each player is independently contributing his "suggested" direction of optimal improvement at point s to the value and gradient of the collective field $\phi(s)$.[15] In such situations the relative weighting to be given to each contribution in determining the gradient at s is of crucial concern. Figure 4.4 parts (a) and (b) represent these two separate components of Harsanyi's tracing procedure for the game of Figure 3.5. The interpretation to be given to the direction of the

Figure 4.2. Potential Field Representing Player 2's Utility Function, $u_2(s)$

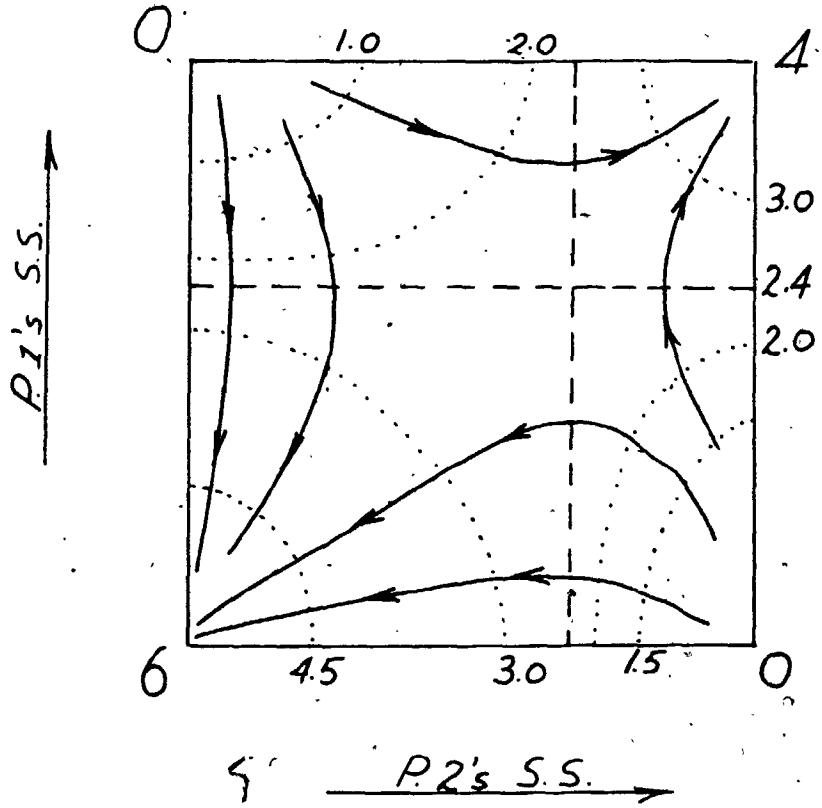

0
1.0
2.0
4
3.0
2.4
2.0
P1's S.S.
6
4.5
3.0
1.5
0
P2's S.S.

Figure 4.3. Potential Field Representing the Product $u_1(s) \times u_2(s)$

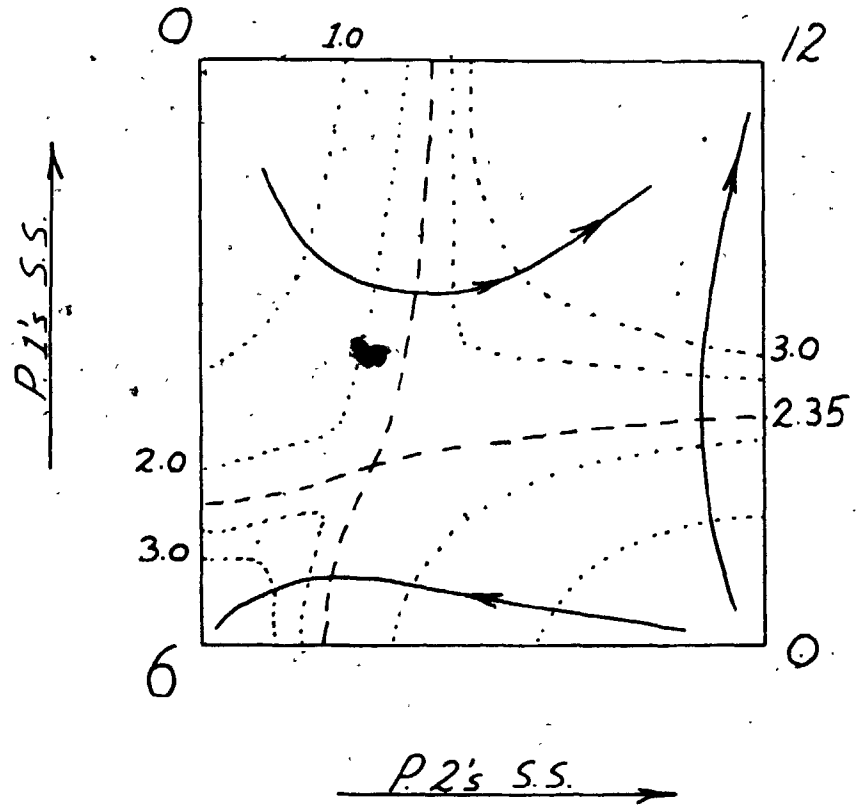
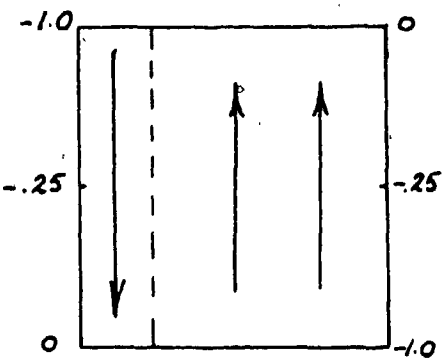
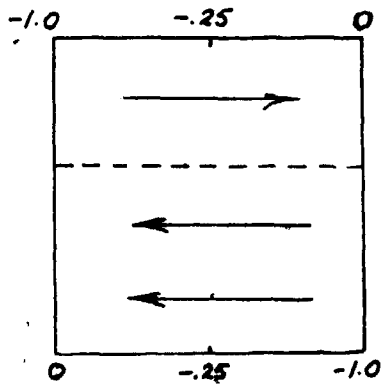

0
1.0
12
P1's S.S.
3.0
2.35
2.0
3.0
6
0
P2's S.S.

Figure 4.4. The Tracing Procedure as a Potential Field

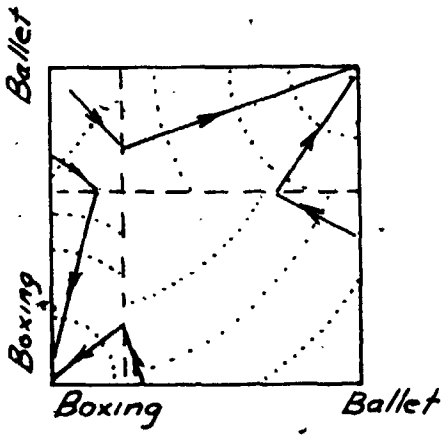
(a)
Husband's
Contribution



(b)
Wife's
Contribution



(c)
Vector Sum
of
Gradients



individually optimal lines of improvement represented in that figure is straightforward. They represent the optimal direction of his own strategy change for each player given the present collective strategy. However, how these individual optimum directions are combined is the real issue. That is, it is in assigning relative magnitudes to these vectors and hence in defining the collective field that the nature of the tracing procedure assumptions becomes apparent.

What the tracing procedure requires is that to each player i's "direction" of optimal change at the point s is assigned a "coefficient" equal to minus the square of the "unitless" orthogonal distance between s and the sub-manifold representing his best reply to $\bar{s}_i$. In the game being discussed this means that the tracing procedure assigns to each player's direction of optimal improvement at $s \in S$, a coefficient equal to minus the square of the probability given by s of his playing his best-reply strategy to $\bar{s}_i$. The vector sum of these "optimal directions" with these coefficients defines the tangent to the line of flux at every point in S, and therefore defines the potential field of the tracing procedure.

Other than its element of symmetry, no straightforward justification for this particular way of weighting the individual optima appears possible. I conclude from this that, considered as an answer to the traditional normative problem of interdependent choice, the estimator $\Pi(t)$ provided

by the tracing procedure is ad hoc and leaves unanswered all the interesting questions.

4.3. Failure of the Psychological Interpretation

The conclusion of the previous section is that the tracing procedure is inadequate as a normative model of practical reasoning in game situations. I now want to argue that for a quite different reason the tracing procedure and any other equilibrium-seeking iterative algorithm based on the naive Bayesian approach to the reciprocal reasoning regress is an unsatisfactory psychological model.[16]

Although the argument is straightforward it deserves detailing if only because it points to a fundamental tension in the use of the normal form of representation as a basis for any psychological model. It has been shown that embodied within the normal form are certain epistemic assumptions which entail a "transparency of reasoning" theorem. One of the important corollaries of this theorem is that only equilibrium points are admissible as solutions to non-cooperative games. However, it must be remembered that this does not entail that the rational-maximizer is an equilibrium-seeker.

Seen as a consequence of the transparency assumption the equilibrium constraint applies only to the joint outcome under specific epistemic constraints, and therefore is only a consideration when two or more maximizers are constrained

as they are by the normal form. Psychologically, the rational player is only a maximizer and for that reason will only adopt a reasoning procedure if it helps improve his chances of a higher payoff. Since this is so the only reason a real-life maximizing player could possibly have for passing on to the next stage in a naive Bayesian regress, let alone all the way to an equilibrium point, is his expectation that by doing so he will improve his payoff.

Now, when an actual player reasons using the regress, I take it this expectation is supported by one particular belief -- the belief that he can reason deeper into the regress than his opponent. In fact at each step of the iteration he adopts the belief that his opponent has stopped reasoning, and is, as it were, a "defenceless non-reasoner" awaiting the devastation of our real-life player's best reply. Of course, the real-life reasoner will then be nagged by doubts, and may very well move on to the next level in the regress. Nevertheless, if at any point that "superiority" belief is not held, and our real-life player believed instead that his opponent was equally capable, equally hard-working, equally perceptive and so on, then we would be hard pressed to justify his entering upon the regress on the grounds of payoff improvement. Of course, if he ignores the regress then so does his equally rational real-life opponent, and the Bayesian iteration never gets off the ground.

As expected, this latter "non-superiority" belief is exactly what the normal form requires each player to hold. Each player knows at the outset that he is not going to "out-think" his opponent since he knows that all players are equally rational. Why then, we must ask, should either of them adopt that mode of reasoning?

The argument here applied to the naive regress can be applied with equal success against the tracing procedure and indeed against any iterative reasoning algorithm with the same overall form as the naive Bayesian model. Whenever the reasoning algorithm is proposed as a psychological model applicable to games in normal form, we discover that the epistemic constraints on that form are so tight that it is impossible to justify the player's taking part in the regress. The "real-life" rational player needs to be allowed to believe that he is smarter, harder-working, or whatever than his opponent, or else he will not reason according to the regress. And that is a psychological constraint Harsanyi's approach to normative issues must learn to live with.

4.4. Conclusion

The past three chapters have discussed Harsanyi's approach to non-cooperative games in some detail. Aside from the obvious inherent value in attempting to clarify the tracing procedure, I have been at pains to draw out the dual nature of Harsanyi's involvement with game theory.

Being an economist, Harsanyi's primary concern is with game theory's potential to provide a foundation for his science. To have game theory perform this function, of course, economics must be viewed as the study of the social consequences of consumer, capitalist and government choices made in interdependent contexts by rational agents. That is, economics must be viewed as general choice-theory. But if this is so, then economics and the rest of the social sciences can be collapsed, and what Harsanyi is attempting to generate is a foundation for the science of social man.

As has been mentioned, if Harsanyi's concern is with the positive use of game theory, he must show how and why his solution concepts are reachable by real-life reasoners. Thus, it is not open to him to respond to the criticisms of this chapter with, "Well, it really does not matter whether this result duplicates conclusions reached by real-life reasoners or not, since my interest is in developing a normative theory". Whatever the value of that stance at the best of times, Harsanyi cannot adopt it without forsaking his positive aspirations.

But if the criticisms of this chapter hold, Harsanyi's dual-interpretation approach has not and may never lead to the desired foundation for social theory. This, of course, is part of what I have tried to bring out. But more needs to be done before that conclusion can be said to be established, and before the alternative approach to the generation of a

normative foundation of social theory, the "procedural" approach, can be seen to be any better off.

The following chapters lend additional support to these conclusions by examining three different approaches to the explanatory notion of social power. The first is Alvin Goldman's non-strategic approach. In analyzing his theory, I hope to demonstrate that a normative foundation is needed for that explanatory concept, since social power depends on the players' ability to generate and employ strategic alternatives. The extension of this conclusion to the claim that therefore, the foundation of social science must be a normative model of man, of course, skips over important issues. Where appropriate, these will be discussed as they arise.

The second approach to social power is Harsanyi's model based on cooperative game theory. What I hope to show in discussing that theory is that for reasons akin to those outlined in this chapter, Harsanyi's cooperative theory fails to support the dual interpretations and does not provide a general explanatory notion of social power. This failure leads once more to the conclusion that game theory and the dual-interpretation approach to rationality cannot provide the sought for foundation for a general theory of social action.

Finally, I examine Emerson's exchange-theoretic approach to social power. In so doing I hope to show, first, that there is a normative underpinning to his version of exchange

theory, and secondly, that that underpinning makes Emerson's approach into a viable theory of "procedural" rationality. The final conclusion is that therefore it is a prime contender as the foundation for the general theory of social action.

FOOTNOTES

1. Nowhere does Harsanyi explicate these criteria. However, he is obviously committed to finding a theory with these properties, as theyare inherent in his dual-interpretation approach to rationality.

2. As before it is for convenience that the normative desiderata are phrased in this "semi-psychological" fashion. The alternative would be to speak impersonally in terms of available data and its optimal use.

3. As has been pointed out, the precise formulation of this criterion is very difficult. The point is that Harsanyi's chosen way of employing normative notions for positive purposes commits him to some psychological criterion along these lines.

4. The following section, section 4.2, examines the most important criticisms of the tracing procedure. In particular, note sub-section 4.2.3 in which the notion of an objective prior is examined, and Harsanyi's response to the criticisms of this section is presented.

5. By calling a prior "non-Bayesian" I mean that it is based on some recognizable alternative to a Bayesian decision rule, for example, an equipartition assumption across pure strategies, or a minimax choice rule.

6. In particular he knows that since his opponents are "rational" they are using the tracing procedure and hence each of <u>them</u> believes of <u>his</u> opponents that <u>they</u> will play according to $\Pi j(t)$, $j \neq i$. Thus each player knows which among each opponent j's best replies he will actually play -- i.e. the best reply to $\Pi j(t)$.

7. Although at first glance the holding of contradictory beliefs might appear unacceptable under <u>all</u> circumstances, it should be noted that the holding of contradictories is only a <u>psychological</u> problem if they are: (a) liable to "arise" and be "employed" in connection with the same problem and (b) susceptible to being conjoined by the agent. If either of these conditions does not obtain, then psychological problems may well be avoided.

8. See Harsanyi's (1975a), Note 3.

9. By a strategy's being "rationally playable" I mean quite simply that there are conditions under which a "rational" player would play that strategy. Whether those conditions obtain is another matter. (Note, for example, that in zero-sum games the minimax strategy is considered to be "rationally playable" even though in general it is a randomized strategy and in the final analysis a pure strategy will be chosen.) In what follows I will assume that the prior represents a mixed strategy for each player, while best replies, the only "rationally playable" strategies once a rational estimate of the opponent's disposition is available, are pure strategies. Both of these assumptions hold "almost always".

10. As will become clear, this notion is intended to help Harsanyi avoid the trap of having the rational player believe both proposition 'p' and proposition 'not-p', and so pertains to both this problem with the priors and to the criticism raised earlier against the use of the estimator $\Pi(t)$. The bulk of his case rests on the tension between what the player "knows" by virtue of knowing the game he is playing, and what he is "rationally" forced into "assuming" because of the players' joint "rational ignorance" about the starting point for the reasoning regress.

11. Harsanyi, (1975a), p.2.

12. In fact, the notion that the prior is some combination of best replies to an assumed distribution (never mind what) makes it look very much like the first step in the naive Bayesian regress. If we were to ask where that assumed distribution came from, and were told that it, too, was a set of best replies to another distribution, which in turn was another set of best replies, and so on, we would be entering upon an infinite backwards regress of the naive Bayesian sort. That Harsanyi needs to avoid this problem is obvious. Just as in resolving the forwards reciprocal-reasoning regress, however, there are only two options open to him: Either he must make the algorithm defining the "starting point" non-Bayesian (i.e. introduce a "trick" to cut the regress short) or he must re-design the backwards regress to show that it converges to a unique point of some sort, i.e. the real prior. Harsanyi has chosen the former route, and as I have argued, the result is not acceptable.

13. Chapter III, section 3.2.4.

14. This makes of player 1 the classic, though naive, "hill-climber" from non-linear optimization theory. He can compute the "value" of where he is, and the "best" way to move, but is never sure whether the "local" optimum he zeroes in on is a "global" optimum as well. To make our naive hill-climber more "sophisticated", we might suggest (a) that he try a variety of "starting points" and compare the local maxima each leads him to, or (b) that he build a mathematical model of the "terrain" and perform a global optimization in analysis. Of course the latter may be impossible in practice, especially when the game involves a number of people and what options they have available alters as the analysis progresses. Under such conditions, then, the ignorance model is not far off the mark. Note 1: One of the crucial problems for the real-life game player is to discover pure strategies which can actually be "played" to give him the information he needs in order to choose the next more nearly optimal strategy. In this sense, all strategies have to be pure for the hill-climbing model to work, for it is only in playing the strategy (as in a "sub-game") that the information required for the next step is obtained. Therefore, the notion of a strategy space needs to be revised for the model to apply in practice. Note 2: This model provides a rich foundation for the dynamical representation of the bargaining process, where the overall game is "played" through the sequential playing of such "artificial" strategies in "sub-games". Unlike Harsanyi's model of the bargaining process (see Chapter VI) this approach allows for these well recognized games of "information" exchange.

15. The "gradient" of a scaler field $\phi(s)$ at any given point in the space is the vector representing the direction of maximal slope at that point. It is written $\nabla\phi(s)$ where the operator ∇ in n-dimensional coordinates is

$$\left(\frac{\partial}{\partial x_1}\bar{u}_1 + \frac{\partial}{\partial x_2}\bar{u}_2 + \ldots + \frac{\partial}{\partial x_n}\bar{u}_n\right)$$

where $\bar{u}_i$ is the i-th unit vector in n-space.

16. As remarked earlier, Harsanyi's Psychological Interpretability criterion is, at best, rather vague. In this section I am merely questioning whether the tracing procedure satisfies the most basic pre-systematic desiderata of a model for psychological processes, and therefore whether the use of Harsanyi's theory for positive purposes has any validity.

CHAPTER V

GOLDMAN'S ANALYSIS OF SOCIAL POWER

5.1. Introduction

Alvin Goldman has recently blown the dust of more than thirty years of philosophical neglect off a conceptual issue which has enthralled philosophers from Aristotle down to Russell: What is it that constitutes social power? With careful analysis he has cleared away some of the rubble plaguing earlier treatments of that question, and has contributed a new theoretical perspective from which the issue may be approached.[1]

Whereas philosophy has neglected social power in the recent past, other disciplines have not. The list of contributions to the discussion surrounding social power, its theoretical foundations, and its empirical manifestations cover the whole of the social and behavioral sciences. John Harsanyi, of course, has contributed his game-theoretic analysis;[2] the social psychologists Thibaut, Kelly, Peter Blau, George Homans, and Richard Emerson have developed their own versions and analyses;[3] Talcott Parsons has been very concerned with social power and authority in his unique social action theory;[4] Shapley, Shubik, and William Riker have proposed concepts and measures appropriate to political

situations;[5] French, Raven, and Cartwright have developed differing field-theoretic approaches;[6] H. A. Simon and James March have elaborated a decision-theoretic model;[7] Robert Dahl[8] has contributed a very influential conceptual analysis. The list of contributions to the discussion goes on with less familiar names, more recent work, and several attempts to put the growing corpus into some sort of perspective.[9] All of this effort has been expended in an attempt to come to grips with the confusions underlying one of the most common and common sensical theoretical notions in all of social science.[10]

In this chapter I propose to examine Goldman's recent discussion of social power and use it as an introduction to a discussion of the strategic nature of power. More particularly, I want to proceed from this discussion of Goldman to an examination in Chapter VI of Harsanyi's game-theoretic analysis, for I believe the most serious drawback to Goldman's analysis is its neglect of the dependence of social power on strategic planning.

My underlying purpose, however, is not just to criticize the proposals of Goldman and Harsanyi, but to use the problems raised by these criticisms to support a thesis regarding what the fundamental conceptual difficulty with that notion really amounts to. In short, I will argue that a strategic, i.e. normative, foundation is required, but game theory and the dual-interpretation approach are inadequate.

5.2. An Action-Theoretic Analysis

5.2.1. Introduction

Goldman's treatment of social power may be succinctly characterized as a deterministic outcome-oriented approach based on counterfactual preference assignments. The closest antecedent to this program is the Shapley-Shubik analysis of *a priori* voting power in committees.[11] In both cases behavior and motivations are pared to extreme simplicity, and clear-cut, outcome-determining decision procedures remove much of the ambiguity surrounding less formal social situations.

However, Goldman attempts a broader analysis of power than the Shapley-Shubik model supported,[12] and he clearly views his work as having laid some sort of foundation for a general theory of social power and social action, which the earlier study never attempted. It is for this reason and because it is the most careful philosophical analysis of power to appear in recent years, that we should not lightly dismiss his program as merely another variation in the Shapley-Shubik mold. Moreover, unlike almost all other discussions in the literature, Goldman tries to develop an analysis of power from an action-theoretic foundation, rather than some special behavioral science perspective.[13] Finally, the failure of his model to support a general notion of power proves to be very instructive insofar as it points clearly at the need for a fundamental revision in our thinking about such matters.

In this section, I present Goldman's basic analysis of power as clearly as possible, then his proposed extension of the Shapley-Shubik measure. At that point I will present my criticisms of his program.

5.2.2. The Analysis

The main objective of Goldman's first paper was to generate an analysis of, and a measure for the amount of, an agent's power over an issue under conditions involving costs and conflict with other agents. This section will present the basic analysis as developed in that paper, and show how it left Goldman with something of a dead end on the problem of developing a general measure of social power.

The first step in that analysis was a treatment of the necessary and sufficient conditions for the attribution of power over an issue. What Goldman proposes is that to possess power with respect to (henceforth "wrt") an issue, an agent must be capable of taking appropriate action so as to ensure the occurrence of at least two outcomes from at least one "partition" of that issue into outcomes.[14] (The requirement of two outcomes is necessary since the status quo is taken to be ensurable by an agent's taking no action at all.)

The capacity to ensure outcomes must be possessed by the agent in the sense that if he should want a particular outcome to occur, then he would enact a series of basic act-types which ensures that outcome's obtaining.[15] It is neither necessary nor sufficient that the agent wants the outcome

which actually obtains. Even if that were the case, to have power he must also possess the requisite resources to have ensured some other outcome should he have (counterfactually) desired its occurrence.

Somewhat more formally, Goldman's analysis works out to be as follows (deleting time indices):[16]

> Agent S possesses some power wrt the issue E iff there exists at least one admissible outcome-partition (P) of the issue (E), with e_i and e_j members of P, and there exists sequences of basic act-types for agent S (σ_i and σ_j) such that:
>
> 1) If S wanted e_i, then he would perform σ_i
> 2) If S performed σ_i, e_i would occur
> 3) If S wanted e_j $(i \neq j)$ he would perform σ_j
> 4) If S performed σ_j then e_j would occur.

The first thing to note about this analysis is that Goldman does not want these four conditions to be read as causal subjunctive conditionals. Rather, they are merely ordinary subjunctive conditionals which he wants to interpret in the following way:

> The truth value of the conditional is the truth value of the consequent in the actual world if the antecedent is true in the actual world, but is, otherwise, the truth value of the consequent in that possible world differing "minimally" from the actual world in which:
>
> 1) the antecedent is true;
> 2) the agent's initial resources, initial beliefs, and initial set of basic act-types remain unchanged; and
> 3) the laws of the actual world obtain.

With this interpretation in mind three additional things should be noted about Goldman's analysis. First, it does not require that to possess power an agent have any influence over the course of events in the actual world. It might be the case, for example, that the outcome a powerful agent wants is not any of those his resources allow him to ensure, but rather some alternative wrt which he might be impotent or even counterpotent. If this were the case, Goldman would want to claim that the agent possesses power since he has potent resources and despite the fact he does not wish to use them.

Secondly, note that conditions 1 and 3 impose epistemic constraints on the possession of power.[18] Goldman's argument is that to possess power an agent must know how and when to wield his resources. More particularly, it must be the case that in the possible world under consideration, that is the one in which the agent desires e_i and it just so happens that if he performed the sequence of basic acts σ_i then e_i would obtain:

1) The agent must "epistemically favor" performing the first element of σ_i as a means to realize e_i.[19]

2) At each point in time thereafter and before e_i has occurred, the agent must at that time "epistemically favor" performing the requisite element of σ_i.

Finally, note that the entire analysis is deterministic. Not only is the agent's "epistemic leaning" required to be

towards a specific sequence of basic act-types (or towards some set of equally good sequences), but the occurrence of the desired outcome must be ensured by his performance, not merely made more probable.

This, then, is Goldman's basic analysis of social power. Note that two concepts one would think are relevant to social behavior -- costs and conflicts -- play no role so far in the discussion. Of course, Goldman does eventually consider how these factors affect his analysis, but it is very revealing that he chose to begin his studies with them in the background.[20] The essential features of how he attempts to bring them back into the analysis will now be presented together in the form of an example.

Consider two agents Row and Column (R and C) each having a set of alternative strategies open to him (that is, each has a set of basic act-type sequences $S_R = \{\rho_i\}$ and $S_C = \{\sigma_j\}$ respectively). Assume that a choice of strategy by each agent fully determines whether or not a particular state of affairs (e), obtains. The "issue", then, is assumed to be the partition {e, not-e}, whose members are the "outcomes".

However, the enactment of a strategy by each of the agents has "other" consequences than the occurrence or non-occurrence of e. These might include the expenditure of resources, incurring penalties, the taking of punitive measures against the other agent, and so on. To compute the "cost" associated by each agent with the "other" consequences, Goldman proposes the following interpretation of the term

"opportunity cost".[21]

First, the "other" consequences of each pairing of strategies (ρ_i, σ_j) is assigned an expected utility by each agent ($U_R(\rho_i, \sigma_j)$ and $U_C(\rho_i, \sigma_j)$ respectively). Secondly, that strategy (say ρ_o) whose "other" consequences have "on balance" the highest expected utility for the agent in question (in this case R) is said to have zero "cost", and this is its cost no matter what act is chosen by C, the other agent. Finally, each agent (again in this case R) assigns every other pairing of strategies (ρ_i, σ_j) a cost value equal to the difference between the expected utility of the "other" consequences for that pairing of strategies and the expected utility of the combination (ρ_o, σ_j). That is:

5.1. $\text{cost}_R(\rho_i, \sigma_j) = U_R(\rho_i, \sigma_j) - U_R(\rho_o, \sigma_j)$

Goldman's example is given in Figure 5.1.

Given this formulation of the interdependence of two agents, we are in a position to ask questions concerning the extent of an agent's power over the issue given costs and the possible conflict of interest between the two agents. Goldman proposes the following approach.

Whether or not an agent will obtain his desired outcome (say e) is, in general, a function of the choice of basic act-type sequences by <u>both</u> agents. But since both agents are assumed to be in some sense "rational", their choices are presumably functions of the "strengths of desire" for the

Figure 5.1. Goldman's Game-Like Matrix

COLUMN (C)

ROW (R)

	σ_1	σ_2	σ_3
ρ_1	0 e 0	-10 not-e 0	-30 e 0
ρ_2	0 e -1000	-20 e -1200	-40 e -1000
ρ_3	0 not-e -600	-10 not-e -500	-30 not-e -600

outcome e relative to the expected costs of the various "strategies" available for obtaining it. Let us assume that R's "strength of desire" for e over not-e is x utiles, and S's corresponding "strength of desire" is y utiles.

Goldman now suggests that given a pair of values (x,y) we should be able to determine from the cost matrix which choice will be made by each agent and therefore which outcome will obtain (i.e. e or not-e). If we map that outcome on a two-dimensional graph as a function of the coordinates (x,y), we will have all the data needed to determine the following power measures:

1) the "extent" (or "amount" or "strength" or "degree") of an agent's power over the issue;

2) each agent's "relative" power over the issue; and

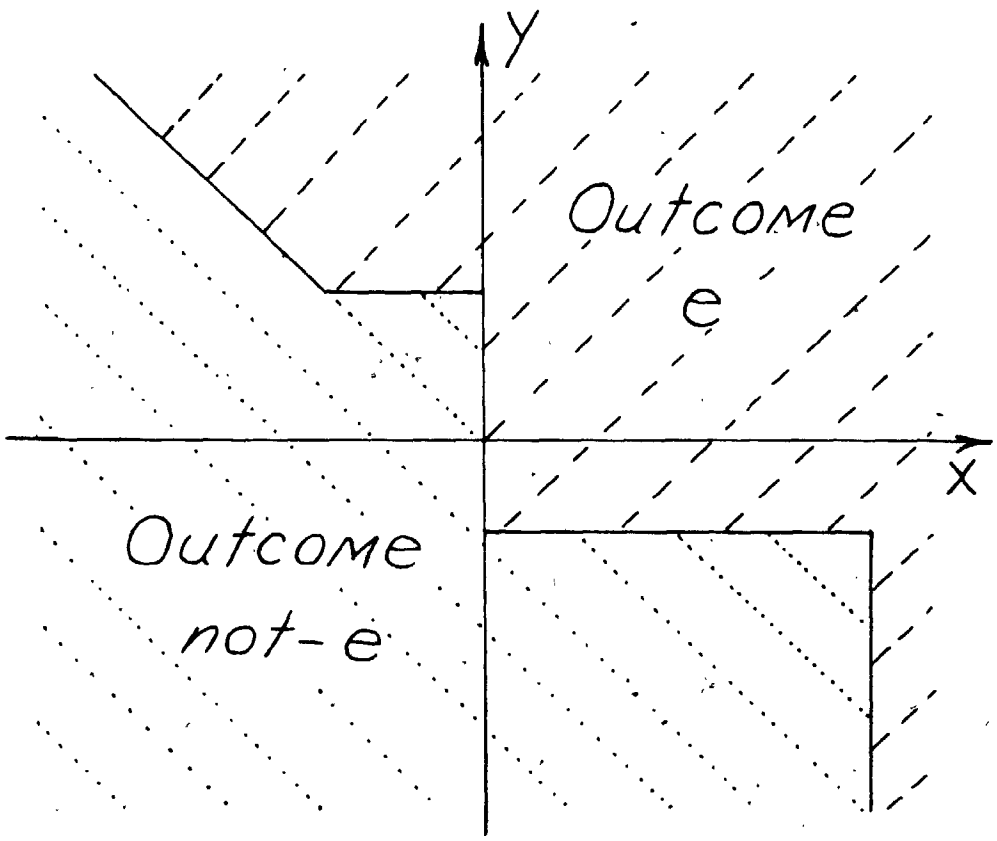
3) the collective power shared by the two agents.

(See Figure 5.2.)

Briefly, Goldman proposes that the extent of an agent's power over the issue {e, not-e} (other things being equal, including the strengths of the opponent's desires) is an inverse function of two relative strengths of desire: that strength of desire (on his part) for e over not-e, hypothetically necessary according to the graph for outcome e to obtain; and the strength of his desire for not-e over e necessary for not-e to obtain. How these are to be combined into a single measure of power over the issue, however, remains an open question.

Relative power is more complicated since under Goldman's

Figure 5.2. An Outcome Graph


Y
X
Outcome
e
Outcome
not-e

basic counterfactual model the strength of desire of both agents must now be counterfactualized. Looking at Figure 5.2 we see that relative power under conditions of conflict of preference must somehow take into account the whole of the NW and SE quadrants of the graph. Goldman recognized that it is not clear, though, whether all that information can be reduced to a convenient non—ad hoc measure. On the other hand it would seem that such a reduction is necessary for his approach to yield any sort of measure of relative power. Unfortunately, Goldman left this issue unresolved.

Finally, collective power is clearly reflected in the shape of the graph in the NE and SW quadrants, that is by the outcome under those combinations of strength of desire corresponding to the two agents having common, not conflicting, interests. Once more, however, it is not clear what measure can be applied to the data summarized in the graph.

At the conclusion of the first paper, then, Goldman appeared to have begun to generate a provocative general program combining a treatment of power with a general account of social action, but had produced no concrete results such as a specific measure function. In his second paper, Goldman re-grouped and adopted a somewhat different tack.[22]

5.2.3. The Measure

In his second paper on social power, Goldman focused specifically on the measurement of relative power over an outcome among some group of agents when whether that outcome

obtains can be made a function of the outcome-wants of the agents in the group wrt an issue containing that outcome. The general structure of his model is the same as Shapley-Shubik's classic model of a priori voting power. In adopting this model Goldman was temporarily discarding the earlier more general approach and once again was ignoring costs and degrees of preference. Nevertheless, he was clearly hoping that this new analysis would inform the dead end he had reached with the earlier framework, to which he could then return.

Goldman's generalization of the voting model is an attempt to satisfy the two usual criteria of adequacy for a measure of an agent's (S) relative power (P_S) over an outcome (e):

1) Ceterus paribus, P_S varies directly as the number of minimally decisive sets (mds) "effective" for e to which S belongs.[25]

2) Ceterus paribus, P_S varies indirectly as the size of each mds to which S belongs.

(Goldman actually adds a third criterion to the effect that S's power is lessened to the extent that the other agents might belong to more or smaller mds, but we may take that for granted in this discussion.)

With this as the basic intuition behind his generalization, Goldman turns once more to the Shapley-Shubik model and borrows their notion of a "pivot". In their model the pivot is the agent whose vote is decisive for the issue in

question when (a) the votes are canvassed in a particular order, and (b) all agents are assumed to be voting the same way on the issue.

Goldman attempts to generalize this notion and considers the pivot to be the agent whose preference wrt the issue (therefore whose choice of behavior) turns out to ensure an outcome's obtaining under the assumption of a particular distribution of preferences wrt the issue across all agents, and given that the preferences of the agents have been canvassed in a particular order. That is, he replaces assumption (b) above with an assumption of a preference distribution across the agents.

With this as the central definition, Goldman proposes a measure of relative power over the outcome (e) which simply counts the number of times an agent's preference wrt the issue would be "pivotal" when all possible "voting" orders and all possible preference distributions are considered. But since it is clearly a mistake to attribute power over some outcome e to a "pivot" when his preference was for not-e, that is when his behavior was supposed to help ensure not-e, Goldman makes the following adjustment to his model. A pivotal agent should be granted credit for power over the outcome that his "vote" ensures in some proportion to his ranking of that outcome in his preference system. If he gets his most preferred outcome, he is a "classic" pivot and gets full credit. The lower the outcome is on his preference

scale, the more "counterpotent" he is, and the less credit he should receive for having been "pivotal".

In the conclusion of the second paper, Goldman admits that his generalization of the "voting" model still remains unrealistic, especially in its failure to account for costs, and in its failure to incorporate any sophisticated treatment of the dependence of behavior on preference orders, especially so as the model does take preference rankings into account in determining power.[24]

We are then left with two Goldmanesque models for social power: one an "unrealistic" modification of the Shapley-Shubik voting model; the other apparently irretrievably restricted in scope to a very special case. Nevertheless, Goldman views his study as being full of promise, and urges that it not lightly be abandoned.

5.3. The Critique

5.3.1. Introduction

There is no question that Goldman ever intended to do anything more than a bit of exploratory analysis of the problems surrounding the notion of social power. His is neither a complete nor a final theory, nor did it pretend to be. Nevertheless, Goldman has made some important suggestions of a fundamental nature and they ought not be passed over without discussion, whatever the superficial lack of results.

In this critique I want to focus my comments on three aspects of Goldman's treatment:

1) his basic analysis of social power;

2) his game-theory-like treatment of costs and interdependence; and

3) his generalization of the Shapley-Shubik model.

I will discuss each in turn.

5.3.2. The Basic Analysis of Individual Power over an Issue

5.3.2.1. Introduction

Probably the most prominent feature of Goldman's proposed analysis of individual power over an issue is its straightforward dependence on the notion of "the ensurability of wanted outcomes". In essence, it is the extent to which an agent could make the world correspond to what he might wish that is taken as the indicator of that agent's power.

If we accept this simple structure as capturing the basic intutition behind Goldman's analysis, there are quite clearly three conceptual cornerstones in the foundation of of his theory:

1) the notion of an "outcome" (and its relation to an "issue");

2) the notion of "wanting"; and

3) the notion of "deterministic ensurability".

In the following sections I first discuss relatively briefly what these three terms might mean for Goldman. I then proceed to an examination of the important problems he faces in

making sense of these key notions in the context of the analysis of individual power over an issue.

5.3.2.2. The analyses of the key concepts

First, concerning the notion of an "issue" and its relationship to the notion of an "outcome" Goldman is quite clearly torn between conflicting intuitions as to which of the two is more primitive. On the one hand, he recognizes that we sometimes regard issues as arising from the conflict between specific alternatives, that is, between competing "outcomes of an issue". In such cases it is the set of "outcomes" which is primitive, and the "issue" is raised over which member of that set will obtain:

On the other hand, he also recognizes that there is appeal in a more general notion of an issue which turns this conceptual dependency around. We sometimes view an outcome as being simply any "resolution" of some "uncertainty" regarding some aspect of the state of the world. Classic examples of an issue in this sense are "the weather" and "the election", where an "outcome" is any more or less specific resolution of the uncertainty attending the issue. For example, "snow" or "fine" might be outcomes of the issue "the weather", while "a liberal won" or "John Lindsay did not win", might be outcomes of the issue "the election".

Of these two intuitions Goldman seems swayed by the former (the "conflict" model) in his second paper, while the latter "uncertainty" approach was the foundation of the

analysis given in the first paper.[25] Which of the two terms is ultimately more fundamental, of course, need not concern us here. But since Goldman appears to be concerned to capture something of both intuitions it will be convenient to develop an analysis which can go either way. I propose the following:

An issue E is a doubleton {W,P} where:

1) W is a set of "possible worlds" each of which is nomologically consistent and is in some sense a "live option" as to how the world is or will be. That is, some action by one or a group of agents will decide which "world" obtains. Note that a "possible world" need not be a complete state description, as our interests will be what determines what "live options" are relevant.

2) P is a set of partitions of W. If P is a singleton Goldman's second treatment is captured, while if P contains all partitions of W the basic idea behind his first paper is captured. However, given his desire to avoid "gimmicky" outcomes, Goldman would probably settle on an intermediate specification of P.[26]

It is important to recognize that this formulation of the issue-outcome relationship implies that <u>two</u> sorts of arbitrary-looking decisions have to be made before the analysis of "power over an issue" can get off the ground. Not only does the issue set get defined without reference to what is of concern to the agent (his preferences are to be counterfactualized and so they provide no help as to what "issue" matters) but a specification of the set of relevant partitions is also required. Goldman hopes that in general there is some "natural" definition of P, but it does not seem likely.[27]

Secondly, on the notion of "wanting" Goldman is relying heavily on the analysis presented in his A Theory of Human Action.[28] Naturally, it is impossible to do justice to that very detailed and sophisticated account of the philosophical problems concerning human action without totally losing our present focus, even though the analysis of action does overlap that of power. Nevertheless, a few comments on that earlier piece are necessary to set the stage for the critique of the concept of wanting as it pertains to social power.

In his earlier discussion of wanting, Goldman can be viewed as facing a dilemma. On the one hand he accepts, even emphasizes, the conceptual and causal linkages between wants, beliefs and actions.[29] On the other hand, he feels that there is no adequate framework or model for treating those concepts as a linked triad. Most particularly, he rejects the rational model for reasons that are never clear.[30]

Goldman's response to this dilemma was to turn away from the challenge of developing a treatment of human action based directly on a model of the conceptual interconnections between wants, beliefs and actions, and to turn to a more traditional analysis treating each term separately with ceterus paribus clauses accounting for the difficult interdependencies. Specifically, this was his approach to the analysis of the notion "wanting".[31]

Now since we often do want to say of an agent that he "wants something ceterus paribus" or that his "overall

attitude towards...(something) is positive", there is some need for such a piece-meal analysis of action-theoretic terms. However, I would want to argue that in general that approach is not adequate. What is required to clarify the issues in action theory is a frontal assault on the three key terms as a conceptually linked triad, and what this entails is taking some sort of rational model seriously. Of course, I will argue for the appropriate model being of the "procedural" sort.

Be that as it may, in the context of social power there is no escaping the connections between the three key terms since those connections are of the essence. In particular, it is folly to adopt at the outset a notion of "wanting" which is independent of social context. It is the dependence of behavior on social context as mediated by wants and beliefs that is to be clarified in the first place. No concept of wanting which ignores the dependence of wants on beliefs, in particular on beliefs regarding the situation at hand, can possibly do the job.

There is a second characteristic of Goldman's treatment of "wanting" pointing directly at one of the serious problems he faces in making that treatment function in the larger context of social interaction. Goldman needs to be able to treat wants as independent variables so that they can be counterfactualized without making other changes in the agent or the world. But in practice wants are not independent in

this sense, even according to Goldman's own model of practical inference.[32] The eventual import of this is that any given "want" can be maintained (or counterfactualized) in many ways.

This ambiguity spells disaster for Goldman's basic analysis of power since he needs to assume that wants alone determine action.[33] But if how a want is counterfactualized affects the subsequent action, it cannot be that wants alone determine action, and so the basic analysis of power falls apart.

There is one final point I want to note about Goldman's handling of the conceptual triad. The only role Goldman assigns to belief in the analysis of power is that an agent must know which basic acts to perform to obtain his desired outcome. The possession or eventual attainment of beliefs serves simply as an "enabling" condition for the agent's performance of requisite acts. The otherwise-noted symmetry between wants and beliefs vis-a-vis action has been destroyed.[34] Goldman wants to say that given a wanted outcome the agent satisfies the epistemic constraints on power simply if he believes the required act is more probable than alternatives to lead to that outcome.

This is perhaps the most revealing limitation in Goldman's treatment of the three key notions of his theory, for it hints at an all-pervasive blind spot. Since the focus of his notion of power is "outcomes", not people, Goldman

never confronts the vagaries of peoples' beliefs and expectations, and so never considers how the interdependence of those beliefs and expectations influences the behavior of all concerned. In Goldman's model either an agent knows the "truth" about the consequences of his acts or he is confused in that regard. No other property of his beliefs is relevant. This treatment of beliefs totally misses the social dimension of social power, and so undercuts the foundation of Goldman's whole program.[35]

This same sort of bias is even more apparent in Goldman's dependence on the third cornerstone of his theory, the notion of "deterministic ensurability". While he does not say that much about this notion, what he does say supports the conclusion that he feels it is necessary to incorporate an extreme form of soft-determinism into the analysis of power. It is not his determinism nor its softness that is at issue, of course, but Goldman's regarding his position on that issue as a necessary part of any acceptable analysis of social power. Specifically what Goldman builds into the analysis of power are: the exclusion of chance events; the assumption that wants uniquely determine behavior; and the conclusion that only those outcomes an agent can ensure by his own actions count toward his having individaul power.

The argument behind Goldman's urging of ensurability is straightforward. If no other agents can affect the state of the world and if there are no chance events, then only

those outcomes which would in fact obtain in the world under consideration could reasonably be relevant to determining the agent's power. The rest -- his hopes, expectations, and so on -- are illusory. An agent has only his ignorance and lack of power to blame if things do not turn out as he wants or expects.

On the other hand, if other agents <u>are</u> involved in deciding the matter, and it happens that only under certain assumptions regarding <u>their</u> behavior that the agent obtains his desired outcome, then he is clearly either totally powerless and at their mercy, or he merely shares "collective" power with them. In neither case does he have "individual" power. Therefore, ensurability is the only characteristic of wanted outcomes relevant to an agent's possession of individual power.

Once again we see the hint of the blindspot regarding the <u>social</u> importance of shared belief. Goldman assumes that an agent's hopes and expectations only affect <u>his own</u> behavior; that is, that they do not affect the behavior of others. Never does Goldman realize that social behavior can turn on the <u>interdependence</u> of (perhaps false) expectations. As a result, he never really gets around to talking about <u>social</u> power.

I now want to turn to a more detailed examination of some of the problems Goldman faces in making these three key concepts supply a footing for his analysis of social power.

In doing so I hope to demonstrate that even if he could avoid these criticisms of his individual analyses, they are still quite unable to function together in the context of social power.

5.3.2.3. Determinism and disjunctions as outcomes

My interpretation of Goldman's position on issues and outcomes left it open whether the set of all partitions of the issue set is relevant to an agent's possession of power, or whether only some subset of it is. I now want to argue two claims on this point. First, that according to his treatment of "wanting" there are really only two options open to Goldman: Either he must consider only that unique partition whose members are singleton sets of the live options (possible worlds) or he must consider all partitions of the issue set. No intermediate position can be maintained, and no other definition of the unique partition is acceptable.

Secondly, I want to argue that according to his notion of "ensurability", only the former definition of P makes any sense. What this entails is that the only sense that can be made of his analysis is as an analysis of "power possessed under an assumption of act-outcome omniscience".

The key to both arguments is the sense we can make of disjunctions of possible worlds within Goldman's theory. So far "outcomes" of an issue can range over both the individual possible worlds in the issue set W and over disjuctions of these possible worlds. The first thing we must now ask is

whether it makes any sense for a measure of power to be a function of the agent's behavior when he "wants a disjunction" but, on the other hand, to be independent of his behavior when he wants the disjuncts separately. The answer, I will argue, depends on the assumed state of ignorance of the agent.

If we assume that the agent knows the consequences of his acts, then it would appear obvious that his power is simply a function of the outcomes of his behavior under the assumption that he "wants" a particular member of W, the relevant set of "live options". Since he knows the consequences of his acts, his behavior whenever he wants a "disjunction" of these possible worlds is the same as his behavior under the assumption that he wants some one of the disjuncts. Discovering his success under the latter sort of assumption, then, is all we need to know to ascertain his power.

On the other hand, if an agent is confused about the consequences of his acts, there is no a priori restriction we can impose on how he might be so confused. For example, he might know enough to be successful in ensuring world "a", but might be so confused that when he wants the disjunction "a v b" he "aims for and misses" the world "b". Since such a blatant confusion is at least possible, if we are to ascertain an agent's "exercisable" power over the issue we are forced to consider his wanting of each of the possible disjunctions of worlds in addition to his wanting of each of the worlds separately. Otherwise the relative degrees of

power of two agents with differing confusions become a function of our selection of which disjuncts are to be considered, and that would not do.

The conclusion is simply that considering only individual worlds validly measures power iff we make the assumption of act-outcome omniscience, whereas allowing the full-range to an agent's confusion demands that we also examine his behavior under the assumption that he wants each possible disjunction of worlds. We might call the former a measure of the agents "possessed" power, and the latter his "epistemically-enabled" power.

However, "outcomes" must be the sorts of things that are "ensurable" as well as being possible objects of wants. On this score, disjunctions of possible worlds do not fare so well.

While "(p v q)" is surely a logical consequence of "(p)", I think it is clear that for Goldman the statement:

1) Action "a" deterministically ensures outcome (p v q)

cannot be a logical consequence of the statement:

2) Action "a" deterministically ensures outcome (p).

Given Goldman's determinism and his approach to possible worlds as "live options" given each act, statement 1) can best be viewed as an elliptical way of saying:

3) Either action "a" deterministically ensures outcome (p) or action "a" deterministically ensures outcome (q), and I do not know which it is.

The importance of the qualification, of course, is to

emphasize that the "or" used in the translation represents the observer's ignorance on the matter, and not any causal or semantical ambiguity.

But this analysis clearly means that statement 3) is not a consequence of statement 2). That is, although we might carelessly speak of "deterministically ensured disjunctions of possible worlds", upon analysis we find epistemological content built into the meaning of the phrase which bans such locutions from functioning in our analysis of power.[36]

Now, if disjunctions are inadmissible as deterministically ensurable outcomes, the only relevant partition of the issue set is the partition having singleton worlds as members. But as we have seen, this partition can only measure "power possessed" under the assumption of act-outcome omniscience. By Goldman's own requirements, then, his analysis can only apply under such restrictive circumstances.

5.3.2.4. On counterfactualizing wants

The central feature of Goldman's analysis of power is its reflection of the capacity of an agent to act and affect the state of the world in such a way as to achieve his goals. Measuring the overall "power" of an agent, then, seems to be a simple matter of determining which goals (among the many goals an agent might choose to pursue) are attainable given his resources and the nature of the world. So far I have argued, in effect, that there is no clearly

unproblematic way to combine this "list" of attainable goals into a single measure of power. Now I want to suggest that this formulation of the problem of social power faces an even more immediate obstacle.

I want to argue that given Goldman's model of intentional action, it is not possible to arrive at even a single unequivocal answer regarding the attainability of a particular goal when one of the antecedent conditions is the counterfactualization of a "want". That is, we cannot counterfactualize "wants" in the way Goldman suggests, feed the result of this operation into his model of action, and get an unequivocal answer regarding the agent's success. Therefore, his analysis of social power does not even get off the ground.

The main thrust of my argument is that the specification of an agent's "want" is not sufficient to determine his choice of action, and is therefore not sufficient to determine whether or not the agent can (given the way the world is) achieve the object of his want.

For the sake of discussion, consider a simple choice situation in which no aspects of the state of the world other than the outcome of the issue are of any concern to the agent. Also consider an issue with three possible outcomes (O_1,O_2,O_3) and an agent facing an uncertain choice in that the consequences of his choice are a function of which state obtains among some set of possible states of the world (S_1,S_2,S_3). Clearly the

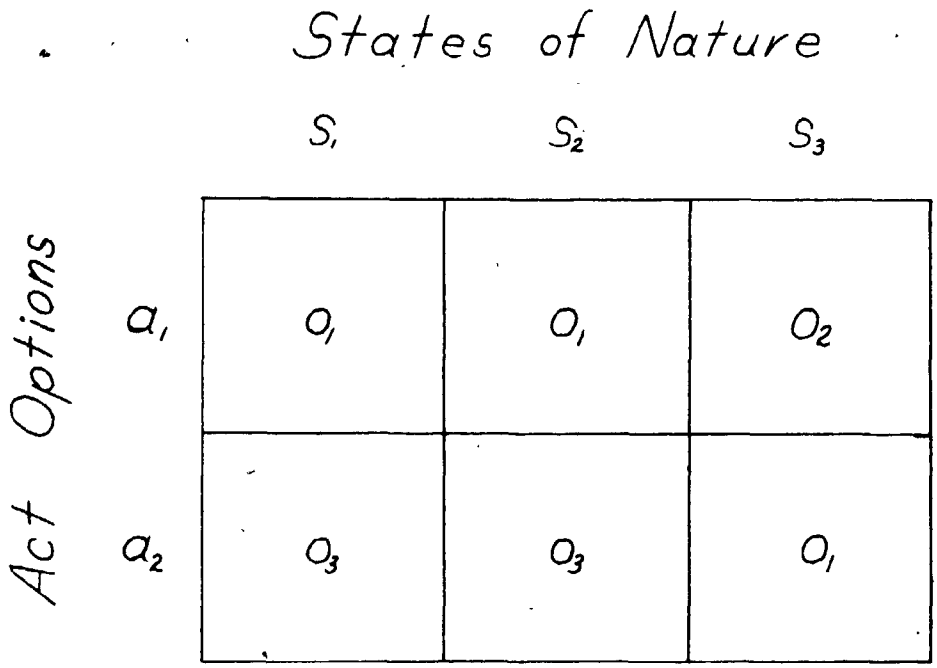
stipulation that the agent "wants" one of the outcomes more than the others is not in general sufficient to determine his choice. The decision matrix of Figure 5.3 illustrates this. In this example, if the agent wants O_1, and the probabilities assigned to the states are (.1, .1, .8), whether he chooses a_1 or a_2 is a function of his ranking of O_2 and O_3 and their relation to O_1.

Another quite distinct problem arises in counterfactual situations. It would seem that Goldman's causal model of practical inference (whereby not one but many of an agent's wants and beliefs enter into the logical inference ending in his performance of an action and where the logical inference is paralleled by a causal chain) requires that in general any given "want" merely occupies some spot in a logical (causal) nexus of inferences. It is a conclusion of some inferences, a premise in others. But if this is faithful to Goldman's analysis of practical inference, it is then not at all clear what the inferences of the hypothetical agent with a "counterfactualized want" would be like. Nor is it clear how nomological consistency can be maintained in a world with such a counterfactual element in a causal chain.

On the one hand, we might simply ignore this problem of maintaining the logical and causal consistency of the agent's inferences, and only concern ourselves with those inferences having the want at issue as a premise.

But even so there is no guarantee that the ignored

Figure 5.3. A Sample Decision Problem


States of Nature
S1
S2
S3
Act Options
a1
a2
O1
O1
O2
O3
O3
O1

logical inconsistency will not (at some later stage in the hypothetical agent's reflection on the problem) come back to haunt him. This is most clearly true in more complex inferences when it is the compatibility of a set of wants that eventually results in action. Moreover, this response does not answer the problem of nomological consistency.

On the other hand, we might take another route altogether and ask the agent what he would do if he had some different "want" vis-a-vis the issue, and leave the maintenance of logical consistency up to him. However, in answering such hypothetical questions few agents bother to exercise the care they would if in fact they were forced to revise some want. Thus, the agent's response can only be a guide to the sorts of revisions that he might undertake, and therefore to his behavior under the counterfactual circumstances.

Finally, of course, we should recognize that part of the difficulty of this problem arises because the revision of a "want" can be carried out in a number of ways. However, if we do turn to something like, for example, a preference ordering, it is not at all clear what Goldman's approach to power would have us counterfactualize when we are to counterfactualize a given want. There are any number of ways to change an ordering to counterfactualize some specific "want", and so in general such counterfactualization will not be determinate.[37] On the other hand, once we start considering general preference relation changes (as opposed to mere

want changes) the whole point of the operation appears to have been lost, since we have moved away from Goldman's basic analysis framework. Moreover, how such preference relations would fit into Goldman's causal analysis of inference is not at all clear.

The conclusion is, I think, quite clear. Goldman needs a more structured preference concept than "wanting" even before he introduces "costs" and "other consequences". But if he were to turn to a rational model and utility theory, the whole foundation of his counterfactual approach to power would crumble, and his causal approach to practical inference and action theory would need drastic revision.

5.3.2.5. Conclusion

In criticizing Goldman's fundamental analysis of individual power I have raised two sorts of problems. The first are problems he faces in clarifying the key notions "issue", "want" and "ensurability". The second are problems with these notions arising in the context of his theory of social power. I focused on four of this latter sort: the effect of determinism on the characterization of an "issue"; the effect of determinism on what sorts of "wants" are allowable; the effect of counterfactualization over wants on the consistency of the agent's system of practical inferences; and the ambiguity inherent in the very notion of a "counterfactualized want".

The overall conclusions I wanted to draw from these

discussions are:

1) We must distinguish more clearly between power possessed and exercisable power, and Goldman's determinism implies that his analysis can only apply to power possessed under conditions of act-outcome omniscience.

2) The counterfactualization of a want without the adjustment of an agent's other preferences or beliefs is contrary to Goldman's characterization of practical inference and the systematic nature of an agent's cognitive make-up.

3) Even if Goldman adjusts his analysis of power to include all the agent's preferences in the determining of action, counterfactualization in terms of "wants" is indeterminate. On the other hand, there is no clear alternative foundation for counterfactualization leading to a clear concept of power over an issue.

5.3.3. The Game-Style Analysis of Costs and Conflict

5.3.3.1. Introduction

Goldman makes two sets of assumptions in introducing his matrix analysis of costs and interdependency. The first set comprises an acceptance of a particular model of rational action in interdependent situations. I want to argue that this model is faulty. The second set comprises an illicit objectification of the distinction between costs and objectives. The result of these assumptions is a totally wrong-headed attempt to introduee a "social" dimension into his underlying model of power.

5.3.3.2. The rationality assumptions

Goldman's analysis of costs and conflict requires

the following as assumptions:

1) There is a rational choice for any agent in any game-like situation.

2) The rational "strategy" is always a "pure" strategy.

3) It is possible to compute an "expected cost" for each act alternative open to the agent.

Without the first and the second, Goldman's determinism falters; without the third his notion of "opportunity costs" does not get off the ground. I want to argue that none of these is well founded.

As to the existence of a rational choice in any game situation, there is surely sufficient disagreement in the literature to at least question whether any clarity has been gained by reducing the problem of social power to that of generating a complete theory of games. Suffice it to say that the connection between interdependent rationality and social power is so well noted that making the blyth assumption that one of the other issue has been resolved adds nothing to either discussion.[38]

As to his assumption to the effect that the rational strategy is "pure", we are faced once more with Goldman's conviction that power ought to reflect what happens in the world and not what the agent expects to happen. Of course, unless each agent performs some specific act, what happens in the world cannot be computed. The choice of a randomized strategy, then, is not sufficient for Goldman.

This clearly runs counter to most recent work in

decision theory, and makes the assumption that a rational strategy exists in every game even more problematic.[39] More importantly, I think it points to an as yet unresolved undercurrent in Goldman's program. This is the connection between his assumption of determinism and the question as to whether he is even talking about social power.

The key point to the following argument is that the important environment for social behavior is that provided by the behavior of other agents, not that provided by an independent nature. Now while it may be appropriate to restrict our attention to "ensurable" outcomes when an agent confronts the forces of nature, when his "opponent" is another equally rational agent another approach is clearly called for. One of the reasons for this is that the expectations of an "opposing" agent, whether in error or not, constitute part of the very real social environment for any actor. Another reason is that, as we have already suggested, rational behavior is often randomized behavior, and what makes the behavior rational is not what will happen, but what is expected to happen.

What these points suggest is that one of the key elements in the analysis of social power must be the role played by belief, in particular beliefs shared by the agents. Since Goldman relegates belief to two rather menial roles (on the one hand that of an "enabling condition", on the other that of a "resource") it is unlikely that he will even

confront this central dimension to the problem of social power.

The third assumption Goldman makes about the nature of interdependent rationality is that the agent will be able to compute an "expected cost" for each act from simply looking at the cost matrix. I want to argue that this assumption begs an important question concerning what the "opposing" agent is going to do.

Since the opposing agent is assumed to be rational, it is clear that what is his best choice is in general a function of what his opponent (the agent under discussion) chooses to do. But the likely behavior of the opponent is information Goldman requires before he can compute an "expected cost", and the computation of an expected cost is part of the procedure by which the rational choice is computed.

Clearly we are dealing with an only slightly altered form of the "reciprocal reasoning problem" discussed in Chapter II. Goldman has apparently not recognized that that problem cannot be resolved by assuming the ability to compute an "expected cost" without begging the question.

In sum, Goldman has introduced an interesting variant of a game matrix in an attempt to inject costs and interdependence into his analysis of power. Unfortunately his use of that matrix requires his making three very questionable assumptions regarding the nature of rational interdependent behavior. The alternative, of course, is to accept

the relatedness of social rationality and social power and attack the problems jointly. Any mere reduction of one to the other simply begs or ignores all the important questions raised by both.

5.3.3.3. On objectives versus costs

In this section I want to explore some of the problems Goldman faces in attempting to develop a rational choice rule around the distinction between the "cost" associated with the "other" consequences of an action and the value attached to desired "outcomes". More particularly, I want to argue the following four points:

1) In his presentation of his choice rule, Goldman appears to want to compute "expected costs" which he cannot do without each agent knowing what the other is going to do.

2) The distinction between "costs" and "objectives" amounts to the assumption of a zero-point in the agent's utility function.

3) The measurement of the agent's preference for one outcome over another in terms of "utiles" imposes an interpersonally meaningful "unit" for the utility functions of rational agents.

4) Points 2 and 3 combine to argue that Goldman's analysis of power once more begs all the important questions.

To compute the "opportunity costs" each agent is to associate with each alternative action, Goldman proposes the following:

> We begin by ignoring all utilities assigned by Row to the outcomes [of the issue at stake]. We next consider the (expected) consequences apart from [the alternative outcomes] themselves, of the various

> alternative sequences of acts open to Row. We then select, among all the alternatives open to Row, the sequence(s) of acts that would yield Row the highest utility. This sequence of acts has zero (opportunity) cost...[40]

The final step is the computation of the opportunity cost to be associated with each pairing of choices (by Row and Column). This is simply the difference between the utility of the "other consequences" of that pairing of acts and the "...(zero) utility of his best alternative".[41]

Now an ambiguity arises directly from this last quotation. Does Goldman mean that the "cost" to be associated with a given pairing is a function of the expected utility associated with the best act given that the opponent's choice is unaltered? (In neither case, of course, is the "utility of the best alternative" zero as Goldman seems to say above. It is only the cost which has been set at zero.)

I want to leave this ambiguity aside, however, for it has already been shown that the approach fails for a more important reason. What Goldman ignores is the obvious fact that in general "other" consequences (and hence their costs) will be a function of both agents' choices. Therefore, to compute "expected" costs will require an estimate of the opponent's likely behavior. But such an estimate cannot be forthcoming until the choice problem is solved. This places Goldman in the tricky position of needing the estimate to get his choice rule to work, but needing the choice rule to get the estimate.

The second of the four points I want to make here is simply to note that the distinction between what counts as an "other" consequence and what counts as part of "the outcome" depends on whether the consequence in question is "welcome".[42] But presumably the notion of a consequence being "welcome" can be translated into utility terms as a "natural" definition of a zero point. I think Goldman would admit to this, and so nothing is to be gained by belaboring it, however much it runs counter to the rational model's employment of revealed preference theory.

However, the third element in Goldman's handling of the interdependent choice problem combines with this assumption of a "natural" zero point to make a mockery of revealed preference theory and leaves Goldman with the problem of developing an alternative if his theory is to have any explanatory value. In determining what the rational "outcome" of an interdependent choice situation is, Goldman assumes that each agent's "degree of preference" for one outcome (e) over the other (not-e) can be assigned an interpersonally meaningful value in "utiles".[43] In doing so, Goldman is assuming an interpersonally meaningful unit of utility. While Goldman may recognize that his approach requires us to assume "interpersonal comparisons of utility",[44] coupled with his assumption of a zero-point for the players' utility functions such a transgression of classical theory leaves his theory with two very obvious "Achilles heels".[45] Neither

assumption can be supported by evidence, and so his whole analysis is untestable. If the notion of social power is to function as an explanatory concept, this is unacceptable.

5.3.4. The Shapley-Shubik Extension

5.3.4.1. Introduction

Goldman's proposed extension of the Shapley-Shubik measure of voting power incorporates four important "changes" to their model:

1) He replaces "votes" and "the passing of bills" with "agents' preferences" and the "outcomes" resulting from the performance of acts by the agents. (Behavior, like voting, is assumed to be a direct consequence of preference.)

2) He drops the assumption that all agents are either unanimously for or unanimously against a bill, and considers all possible distributions of outcome preferences.

3) He extends the model to cover multi-outcome issues.

4) He advances the proposal that a "pivot" be judged to have power in some proportion to his ranking of the outcome his action ensures.

The central thrust of my criticism is that of these four changes 2) shows most clearly that Goldman has misunderstood the reasoning lying behind the Shapley-Shubik measure. His commitments to the notion of "wanting" and to counterfactualization have in this instance led him seriously astray.

With respect to his other three proposals, his misunderstanding of the role of coalitions in the Shapley-Shubik model makes his generalization unable to capture the nature

of the strategic possibilities in the n-person interdependent context of concern. Once more Goldman can be seen to fall into error for underestimating the link between power and rationality.

Simply stated, the Shapley-Shubik measure was proposed as a "method for evaluating a priori voting strength". The resources of each agent are simply the number of votes he controls, as costs, incentives, log rolling etc. are not involved a priori. The situation is pristine -- votes determine outcomes according to a given rule and the number of votes each agent controls is given data.[46]

Clearly, Goldman would have had three reasons for choosing this sort of model as the basis for his measure of power:

1) The approach measures power prior to the intrusion of most empirical factors like party allegiance and costs, just as Goldman feels power ought to reflect an agent's ability to succeed independently of such empirical factors as "what he happens to want". In short, both derive power from consideration of counterfactual circumstances.

2) The voters have a single simple goal, just as Goldman's basic model of power has agents "wanting" outcomes.

3) The voting model has a clear-cut two-way decision procedure -- either the bill passes or is blocked -- just as Goldman's deterministic and omniscient approach to the consequences of an act entails either the agent's succeeding or failing in his attempt to satisfy his want. Intermediate results are irrelevant in both cases.

These three points correlate each of the cornerstones of Goldman's basic model with an obvious counterpart in the Shapley-Shubik model. To Goldman the extension of their measure to a general behavior model in keeping with his basic approach must have appeared to be straightforward. I now will argue that that hope was founded on a crucial confusion.

5.3.4.2. Pivothood and coalitions

The Shapley-Shubik measure makes an agent's power equal to the prior probability of his vote deciding an issue. In their original paper that probability was computed by considering all possible voting sequences and counting the fraction of times the agent was pivotal under the assumption that all previous agents in the sequence voted the same way. Of course, this was only one way of computing that prior probability, and perhaps it was a confusing way to look at the problem of a priori power -- at least it confused Goldman.

What Goldman failed to realize in proposing his "extension" was that from a strategic point of view an agent's power in the various coalition possibilities are dependent on one another. In taking preferences as primitive, and thereby assuming that each preference distribution is independent of the others, Goldman ignores the strategic dimension of social power. The point he misses is that an agent's power in any particular coalition arises both from his capability as a single individual to withdraw from the

coalition, and from his being a member of many of that coalition's sub-coalitions. In each sub-coalition he also wields power, and therefore must be judged to have some control over the actions taken by that small group as a group.

With respect to the larger coalitions, then, he also has some control over whether each sub-coalition will defect, and so any measure of his total power in that larger group must reflect this fact.[47]

Through taking the preference distribution as primitive and the coalition structure as defined, rather than having the coalition structure variable, Goldman could not accommodate this dependence of the agent's power as a member of, for example, the grand coalition of all members (i.e. where they all vote the same way) on his power as a member of each of the smaller coalitions. He assumed these were independent situations, whereas strategically speaking the power in the grand coalition is based on power in the smaller ones.

To clarify further this dependence of power on strategically possible coalitions will take us into problems I would rather delay until we consider Harsanyi's theory of power in Chapter VI. The important point to make here can be stated simply: Goldman has misunderstood the role of pivothood in the Shapley-Shubik measure. Being pivotal is not important per se, since in the situations being modelled votes or actions are either taken simultaneously or in some

fixed order. The reason Shapley and Shubik employed voting sequences and pivothood was because being pivotal under the special assumptions they made just so happens to reflect the contribution an agent's voting strength makes both as an individual and as a member of the many strategically possible coalitions. Of course, the various possible distributions of preferences wrt the issue as introduced by Goldman, are part of what constitutes these possible coalitions, since "strategically" speaking, an agent may threaten to vote against his most preferred outcome if he has reason to believe that the threat will be taken seriously. To compute a priori power by considering pivothood is only a convenient mathematical trick. A better sense of why this trick works is obtained from an interpretation of their model treating coalition structures directly. In so doing, it quite legitimately disregards the artificial device of pivothood, and concentrates on an agent's strategic alternatives as a source of his power.

In short, an agent's threat in every case is to join the opposing coalition, but he can threaten this both as an individual and in concert with the members of sub-coalitions. In considering both pivothood and alternative preference distributions (i.e. coalition structures) Goldman is guilty of double-counting, and has shown that he completely misreads the importance of coalitions and strategic threats to the Shapley-Shubik measure.

The end result of this confusion is that once a preference distribution is fixed, there is no justification for assigning power on the basis of an agent's being pivotal. By assumption, he has no alternative actions open to him, therefore he has no threat capability, therefore he has no strategic power. In the Shapley-Shubik measure the prior probability of being pivotal does not pertain directly to an agent's power -- it is only of concern when it reflects the prior probability of his vote deciding the issue, while for his votes to "decide an issue" an agent's defection must change the outcome. However, once Goldman assumes a preference distribution for an agent, that agent's defection is not an issue, and so neither is his pivothood.

5.3.4.3. Conclusion

Without a doubt each of the four modifications Goldman suggests is crucial to any generalization of the Shapley-Shubik model. Unfortunately, from the outset his extension is unable to do the job of incorporating the changes because it fails to develop a generalized role for strategic considerations, and it was these that lay behind the original model.

As it happens, there has existed for some time a generalization of the Shapley-Shubik model of power based directly on a generalization of its strategic foundation, i.e. on a modified Shapley value for n-person games. This is Harsanyi's theory of power and n-person cooperative games.

I will now turn to a discussion of that theory and hence pick up the thread of Harsanyi's theoretic analysis of rational behavior.

Of course, it would be absurd to suggest that Goldman's failure to provide an adequate analysis of the explanatory concept of social power on a non-normative foundation by itself means that none could possibly be forthcoming. However, that conclusion is supported by the type of criticism levelled at his model. Some way of handling rational choice in interdependent contexts would seem to be necessary to an analysis of that concept, as would some way of handling notions like "strategic alternatives". It is these considerations which argue for the development of a normative foundation for such explanatory notions.

Of course, there is still room to dispute the importance of the notion of social power in the social sciences, and so to counter the generalization of the above conclusion. Riker, for example, confesses:

> The final question...concerns the appropriate scientific attitude toward the conception of power itself. Ought we to redefine it in a clear way or ought we to banish it altogether. My initial emotion, I confess, is that we ought to banish it.[48]

That this is not my attitude should be clear. The central problems in the social sciences concern how one agent's beliefs, values and actions are related to those of other agents. Whether or not social power is the key notion in understanding these interdependencies, some such notion

will have to be invoked, and my contention is that that notion will have to confront the same obstacles raised here against Goldman's theory. If this is the case, then the conclusion that explanatory notions in the social sciences demand a normative foundation will go through, and the criticisms raised against Goldman will be generalizable.

FOOTNOTES

1. The two key papers are Goldman's "Towards a Theory of Social Power" (1972), and his "On the Measurement of Power" (1974). Russel's contribution is his Power, A New Social Analysis (1938).

2. Harsanyi's main discussions are in his "Measurement of Social Power, Opportunity Costs, and the Theory of Two-Person Bargaining Games" (1962), and his "Measurement of Social Power in n-Person Reciprocal Power Situations" (1962). See Chapter VI for a detailed discussion of his theory.

3. See Thibault and Kelly (1959), Blau (1964), Homans (1974), Emerson (1962) and Emerson (1972b). Emerson's theory of power is discussed in Chapter VII and Chapter IX.

4. See, for example, Buckley's discussion in his (1967), p. 183 and the references.

5. See Shapley (1953), Shapley and Shubik (1954), Riker (1962) and Riker (1964).

6. See French and Raven (1968) and Cartwright (1968).

7. See Simon (1957), and March (1955).

8. See Dahl (1957).

9. See especially Schopler (1965).

10. As Schopler notes in his (1965), p. 177: "In his 1953 presidential address to the Society for the Psychological Study of Social Issues, Darwin Cartwright [contended] that any social psychological theory was incomplete without this construct."

11. See Shapley and Shubik (1954).

12. The Shapley-Shubik model was designed to provide an a priori measure of committee voting strength. There was no attempt made to generalize the treatment to handle other sorts of interactive contexts.

13. As a survey of the literature cited in footnotes 1 to 9 of this chapter shows, many of the recent analyses of social power have assumed some specific social or behavioral model of the causes of human behavior. Goldman has attempted to steer clear of making any assumptions of that nature, with the unfortunate result that his concept of social power seems unable to function in explanatory contexts.

14. A partition of some set S is any collection of subsets of S such that every member of S belongs to one and only one of those subsets. For Goldman the "issue" is the set S and the subsets represent "outcomes". (See below section 5.3.2.2 for a further discussion of this relationship between issues and outcomes.)

15. In his (1970) Goldman goes into considerable detail on the nature of "basic act-types". For the purposes of this discussion of social power, the main import of the term is that a basic act-type "is one that a person can do 'at will', an act-type that is 'directly' under his control", Ibid., p. 225.

16. See Goldman (1972), p. 226.

17. As Goldman points out, this interpretation owes much to Robert Stalnaker. See Stalnaker's "A Theory of Conditionals" (1968).

18. The whole issue of the relationship between belief and social power is a problem for Goldman. The present analysis indicates that he would really rather assume that the agent knows which act sequence among those he has available will give him his wanted outcome. Thus he reduces the holding of "appropriate" beliefs to an enabling condition. See below section 5.3 for a discussion of this problem.

19. For an agent to "epistemically favor" a certain act-sequence means, for Goldman, that he believes that that sequence is more probable than any other to lead to the wanted outcome (1972), p. 230.

20. What this seems to reveal is that Goldman considers the essential property of the concept of social power to be non-social. His whole analysis is constructed on the model of an agent's manipulating nature, rather than on his interacting with other agents. While this line of analysis might give rise to an interesting general concept of "power" (as, for example, in "an agent has

the power to do act x", and "the waterfall has the power to light many houses"), there is no reason to expect that this approach will reveal much about the peculiar social nature of social power.

21. Ibid., 251. In section 5.3.3.3 I argue that it is unclear exactly what Goldman means here, but I think this interpretation captures the gist of his proposal.

22. At the end of the second paper Goldman admitted that there are "serious complications" involved in extending this approach to multi-outcome multi-agent interactions. This may have been what motivated his consideration of the Shapley-Shubik model. See his (1974), p. 252.

23. A set of agents is "effective" for an outcome if it can ensure that outcome against concerted opposition. It is "minimally decisive" if there is no subset of S also effective for that outcome.

24. Ibid., p. 252.

25. In his (1972) he concentrates on examples like "whether or not it rains at a particular time and place" (p. 223) and tolerates many partitions of an issue into outcomes. In (1974) he says, "An outcome is a possible event or state of affairs. An issue is associated with a set of two or more outcomes that are mutually exclusive and jointly exhaustive; different sets of outcomes determine different issues" (p. 232).

26. A "gimmicky" set of outcomes involves outcomes differing in "irrelevant" respects. For example, concerning the issue "the weather", the following is a "gimmicky" partition into outcomes: (It rains with my hat on, It rains with my hat off, It doesn't rain). Goldman wants to avoid attributing power over the weather to an agent who can ensure either of the first two outcomes simply because it happens to be raining and he has the power to take off his hat.

27. Footnote 26 above gives an example illustrating why P cannot contain every partition; but as Goldman recognizes, it is not easy to see how to demarcate "gimmicky" outcomes. (See (1972), p. 266.) On the other hand, allowing only a single set of outcomes to define an issue overlooks obvious comparisons we want to be able to draw. Consider, for example, an agent who can ensure "rain" and one who can ensure only "precipitation". Since these are not mutually exclusive outcomes, under the second proposal they concern different issues and

34. See his discussion of the rational model in (1970), pp. 136-137.

35. Consider, for example, two "street" agents who both believe that one of them has a devastating left hook. This "shared belief" gives that agent great power over the other even though in fact he might not know the first thing about fighting. The shared belief is enough to convey power to him even where there is no "truth" to what they believe.

36. I do not want to suggest by this argument that Goldman's analysis of ensurability demands that each member of W will be a complete state description, for we have already allowed that it is our interests (and ignorance) that determines the issue set. What is at issue here is whether our interests (and ignorance) should be involved after that choice has been made. I have argued that if we answer in the negative, then deterministic ensurability demands that we can only include in P singleton members of W.

37. Consider the agent's preference ordering $O_1 > O_2 > O_3$. Goldman must maintain that the agent "wants" the outcome O_1. But to counterfactualize the want, we could either move to the ordering $O_2 > O_1 > O_3$ or to $O_2 > O_3 > O_1$, and an agent's behavior may depend on which we choose.

38. I am, of course, suggesting two things here. First, as the discussion of Harsanyi's theory of non-cooperative games tried to show, there may be no way to characterize a unique rational solution to many interactions conceived in the usual game-theoretic sense. Secondly, the notion of power is very much a strategic concept, so that although debates in game theory concerning strategic interaction will surely inform the analysis of power, they should not be assumed to have resolved all the difficult problems concerning that concept.

39. Goldman's arguments against a probabilistic treatment of behavior (see his (1972), p. 268), amount to a denial that people might ever decide to "flip a coin", or "roll a die" in deciding what to do. (Either that or else he wants to regard the outcome of those events as being determinate, too.) But this argument clearly runs counter to his assumption of an agent's rationality, as any introduction to game theory shows. (See, for example, Rapoport's (1966), Chapter 6.)

40. See his (1972), p. 250-251.

41. Ibid., p. 25.

42. Goldman recognizes this. He states: "If C would be an unwelcome consequence of e (eg. going to jail), then it is included in the category of cost." Ibid., p. 267.

43. This must be the case for the following to make sense: "Let us assume...that the degree of preference for both agents is the same [my emphasis]". Ibid, p. 252.

44. He does recognize that "It is assumed that interpersonal assignments of utility can be made." Ibid., p. 250.

45. See Luce and Raiffa (1957), p. 33-34 for a discussion of interpersonal comparisons of utility.

46. See, for example, Shubik's Game Theory and Related Approaches to Social Behavior (1964), p. 141.

47. This is the main point to Harsanyi's consideration of all possible syndicates. (See Chapter VI.)

48. Riker's "Some Ambiguities in the Notion of Power" (1964), p. 348.

CHAPTER VI

N-PERSON GAMES AND HARSANYI'S THEORY OF SOCIAL POWER

6.1. Introduction

In an important pair of papers published in 1962, John Harsanyi moved the discussion of social power into the context of his theory of cooperative games.[1] Involved in this shift was a critique of earlier theories (in particular the proposals of Dahl, March and Simon)[2] to the effect that although their emphasis on the behavioral effects of an agent's influence attempt was correct, what they neglected to take into account were the opportunity costs to the two agents of the exchange. That is, if we assume agent A is attempting to influence agent B, what had been excluded from consideration were the cost to agent A of his attempt to influence agent B and the cost to B if he should refuse to yield to A. As Harsanyi put it:

> A's power over B should be defined not merely as an ability by A to get B to do X with a certain probability ρ, but rather as an ability by A to achieve this at a certain cost u to himself, by convincing B that he would have to bear the total cost v if he did not do X.[3]

The major argument Harsanyi offered in favor of incorporating the first of these cost factors was the observation that we would not be inclined to attribute high power to an agent whose attempt to influence the behavior of

another could succeed only at very high cost, for example at the cost of his own life.[4] Of course, if the consequences of a successful influence were important enough to such an agent, he might make the attempt even though the costs were high, but we would nevertheless feel inclined to take the expected costs of the exercise of influence into account in computing the power he possesses.

The other cost factor, the opportunity cost to agent B of non-compliance, must also be considered, Harsanyi notes, if the power concept is to function as an intervening variable in the explanation of B's behavior.[5] If we are to regard B's decision to adopt some policy X to be a function of the advantages and disadvantages he associates with that policy and its alternatives, then the explanation of his decision (using the power concept) must take into account the difference A's intervention has made. This is measured by the strength of B's incentive to adopt policy X, which in turn is a function of the cost he associates with the effects of non-compliance.

Harsanyi next suggested that, in general, in two-agent interactions involving conditional rewards and sanctions, the unilateral power exercised by one of the agents over the other is only half the story.[6] If nothing else, the influencee will surely press for a reduction in sanctions or an increase in rewards by threatening to withhold compliance, even though it may cost him to do so. In so doing he will exercise whatever power his ability to be stubborn gives him over

the influencer.

The general situation, then, is surely one in which both agents to a two-party power relationship attempt to influence the other, and so is one of bilateral, or reciprocal, power. As noted, this avoids the pitfalls of Goldman's approach. More specifically, since in such a situation the behavior of both agents is a function of implicit or explicit bargaining between them, the proper context for examining the notion of social power is the theory of games.[7]

Harsanyi faces three separate problems in developing a general measure for social power through game theory. First, he must have available a defensible solution concept for two-person bargaining games. Secondly, he must develop an n-person solution concept on that foundation, since most power interactions involve more than two agents. Thirdly, these solutions must be shown to generate a measure of social power satisfying at least some of our intuitions regarding that notion. Harsanyi's treatment of these three problems will now be presented. A critique of his program follows in sections 6.7 and 6.8.

6.2. Harsanyi's Bargaining Approach to Two-Person Cooperative Games

The essentials of Harsanyi's approach to two-person cooperative games are straightforward. As usual, we are to regard each agent as having a number of pure strategy options open to him, the consequences of which are a function of the choices by both agents. The agents are endeavoring to agree through bargaining with one another upon some jointly randomized strategy giving each an optimal expected payoff under the constraints of the game. In the event that no agreement can be reached, each player threatens to choose to play some "conflict strategy", chosen so as to demonstrate to his opponent the damage he is able to inflict at low cost to himself.[8]

The choice of this conflict strategy, or threat strategy, and how it determines the solution of the bargaining game are the two central dimensions to Harsanyi's two-person model. On both counts Harsanyi follows John Nash.[9]

The Nash solution to a two-person bargaining game as a function of <u>given</u> conflict payoffs is that jointly randomized strategy which maximizes the product of the two players' increases in expected payoff relative to those conflict payoffs. The Nash criteria which only this solution satisfies are well known.[10] Harsanyi has generated an alternative set of such "rationality" postulates.[11]

Since the choice of conflict strategies has such a direct bearing on the bargaining outcome of the game, each

agent's choice of conflict strategy could reasonably be made with the sole criterion of admissibility being the effect of that choice on his final payoff. This is the Nash-Harsanyi approach. Of course, since each agent's choice is made independently of the other, and yet the effect of each choice on the final payoff is a function of both choices, the condition being sought is one of "mutually optimal" conflict strategies.[12]

Nash, again, has shown that there always exists such a pair of mutually optimal conflict strategies, and that all pairs of mutually optimal conflict strategies in any game will lead to the same final payoffs.[13]

These two Nash results, then, form the foundation of Harsanyi's theory of two-person bargaining games. The mathematics in which this approach to the two-person game is formulated need not concern us here, however, as the n-person situation is what matters more, and the crucial problem there does not arise from the two-person mathematics. Rather, the most pressing issue in Harsanyi's n-person theory is how the larger game can be constituted out of two-person subgames, or how a set of two-person Nash solutions can form equilibrium conditions defining a solution for an n-person interaction.

I will discuss Harsanyi's extension of the Nash approach after a brief introduction to the traditional formulation of n-person games.

6.3. Games in Characteristic Function Form

Much of the analysis of n-person games has been undertaken with a particular set of assumptions being adopted to characterize the game context. These assumptions entail a reduction of the game from normal form, the usual representation in two-person theory, to a representation called the "characteristic function form".[14] These assumptions can be summarized as below:[15]

1) Coalitions of subsets of the set of n players (N) can form freely and pre-play communication is unrestricted.

2) By suitable choices of the zero-point and unit of the utility functions of the n-players the payoffs of the n-person game can be converted to units of "transferable utility", thus allowing side-payments in these units among the players of any coalition.

3) The payoff of concern to any coalition (S) is that which S can guarantee itself; that is, it is the "security level" of S when the n-person game is represented as a two-person game with S being opposed by the counter-coalition N-S (denoted $\bar{S}$).

Together with certain normalization adjustments and the elimination of so-called "inessential" games,[16] these assumptions permit the representation of any n-person game by a set-theoretic measure defined over the set of $S \subseteq N$, i.e. over the set of all coalitions. This measure, the characteristic function v(S), has the following properties:

$$(6.1) \quad \underset{i \in N}{\forall} \; (v(\{i\}) = 0)$$

$$(6.2)\quad v(\emptyset) = 0$$

$$(6.3)\quad \underset{S,T\subseteq N}{\forall} \; (S \cap T = \emptyset \supset v(S \cup T) \geq v(S) + v(T))$$

$$(6.4)\quad v(N) = 1$$

Many solution concepts have been proposed for games in characteristic function form. Some constitute a set of payoff vectors in the form $x = (x_1, x_2 \ldots, x_n)$ to which rational players should restrict themselves. One of the most basic of these, the "core", is based on the assumption of both individual and collective rationality.[17]

All players are individually rational if for any vector x in the solution:

$$(6.5)\quad \underset{i\in N}{\forall} \; (x_i \geq v(\{i\}))$$

or, for the normalized game:

$$(6.6)\quad \underset{i\in N}{\forall} \; (x_i \geq 0)$$

The players are collectively rational if no subset of players will be satisfied with getting less than they could by joining a coalition. That is:

$$(6.7)\quad \underset{S\subseteq N}{\forall} \; v(S) \leq \sum_{i\in S} (x_i)$$

In particular, from 6.7 it follows that:

$$(6.8)\quad v(N) = \sum_{i\in N} (x_i)$$

or, for the normalized game:

(6.9) $v(N) = 1$

Now, these constraints on the admissible payoff vectors constitute the definition of the solution concept called the "core". As it happens, however, some games do not have a core, and this concept, though simple enough, cannot really be accepted as defining a general solution for all games in characteristic function form.[18]

The solution concept for games in characteristic function form most closely linked to Harsanyi's and to the search for a power concept is the "Shapley Value Solution".[19] The assumptions lying behind this approach are as follows:

1) The grand coalition of all players will eventually form, since that coalition maximizes the payoff available to the players.

2) The order in which the players enter the coalition is random.[20]

3) Given some specific order of formation, the amount each player i receives from the "value" of the grand coalition equals the amount of transferable utility his joining the already formed group adds to its value.

i.e. $\phi_{\delta}(i) = v(S) - v(S-\{i\})$

where $\phi_{\delta}(i)$ = the payoff to player i given this order of formation (δ)

S = the coalition after i joins

$v(S)$ = the characteristic function for this game

With these assumptions Shapley shows that the "value" to player i of the n-person game represented by the

characteristic function v(S) is given by:

$$(6.10)\quad \phi(i) = \sum_{S \subseteq N} \frac{(s-1)!(n-s)!}{n!}[v(S)-v(S-\{i\})]$$

where $s = |S|$

$\phi(i)$ = value of the game to player i

6.4. Problems with the Shapley Value and the Characteristic Function Form

A number of objections have been raised against the characteristic function form of representation. Luce and Raiffa,[21] for example, point out with the aid of an example taken from McKinsey,[22] that the characteristic function abstracts away asymmetries found in the normal form, asymmetries a player of the normal form might use to his advantage in bargaining with his opponents. They also suggest that the implicit use of security levels to represent the expected value of coalitions is probably inadequate in the case of non-constant sum games.[23] In such cases, opposition by the counter-coalition raises the possibility of bargaining and therefore of expected payoffs exceeding the (very conservative) security level.

In exploring this idea further, Harsanyi argued that the threat potentials of the opposing coalitions ought to be used to define the value of a coalition instead of the security levels.[24] He introduced a "modified" characteristic function into the Shapley definition of value to reflect this, and, as we shall see, this modified Shapley Value turns out to

this was also a return to extremely difficult technical problems.

6.5. Harsanyi's Theory of N-Person Cooperative Games

6.5.1. Introduction

Harsanyi's bargaining approach seeks to define a solution for the n-person game through examining the possible equilibrium conditions existing among the solutions to a set of bargaining sub-games. These sub-games involve subsets of the n-players called "syndicates". (Unlike coalitions, syndicates never actually form, even in theory. Rather, they represent the interdependencies among subsets of the n players which rational players must take into account in determining their bargaining strength in the original n-person interaction.)

The crux of Harsanyi's model is that these syndicates allow the players to determine how much power they would wield if they were to act in concert with others. The model Harsanyi proposes empowers each syndicate to do the following:[28]

1) It announces a threat strategy to which the members of the syndicate are bound in the event of conflict with the counter-syndicate.

2) It announces dividends to its members (non-transferable utility payoffs) which all members agree to secure through the cooperative choice of threat strategy.

Since these dividends are to be considered additive, that is, a player's final payoff is simply the sum of his dividends, and since Harsanyi is seeking equilibrium conditions, his model reduces to considering what each syndicate's threat strategy and dividend guarantee are given the threat strategies and dividend guarantees of all other syndicates.

It should be noted that this bargaining model involves two important modifications of the Shapley approach. First, the characteristic function's use of the security levels of the coalitions in defining the outcomes of a confrontation with the counter-coalition is replaced by Harsanyi's use of conflict payoffs based on each syndicate's bargaining strength or threat potential. This is equivalent to Nash's general bargaining solution to two-person cooperative games.[29]

Secondly, Harsanyi has been able to drop the assumption of transferable utilities and to generate a solution concept having the modified Shapley value as a special case.[30]

6.5.2. Definitions and Assumptions

The fundamental problem confronting Harsanyi in developing this generalization of the Nash approach was that of generating sufficient equations constraining the solution as the number of players increased, but to do so without introducing more variables than his constraints can handle. To see this, consider the basic Nash approach for n-person

games with known disagreement vector, that is, when the payoffs to the players in the event that they cannot agree on a joint strategy is given by the rules of the game. The following equations define a unique solution vector ($\bar{u}$) assigning a payoff to each of the n players:[31]

$$(6.11)\quad H(\bar{u}) = 0$$

$$(6.12)\quad H_i(\bar{u}) = a_i \quad \text{for all } i \in N$$

$$(6.13)\quad a_i(u_i - d_i) = a_j(u_j - d_j) \quad \text{for all } i, j \in N$$

where $P = \{X\}$ = the prospect for the game Γ

$X = (X_1, X_2, \ldots, X_n)$ = some member of P

$d = (d_1, d_2, \ldots, d_n)$ = the disagreement vector given by the rules of the game

$\bar{u} = (u_1, u_2, \ldots, u_n)$ = the solution vector

$H(X) = 0$ is the equation of the upper right boundary of P (it is the set of points in P undominated by any other points in P)

$H_i(X) = \frac{\partial H}{\partial X_i}$ = the first partial derivative of H with respect to X_i

When the disagreement vector is given by the rules of the game, this system of 2n equations (there are only (n-1) independent equations of the form 6.13) contains only 2n unknowns, and can be solved to yield the unique solution ($\bar{u}$) given by the equation:[32]

$$(6.14)\quad \bar{u} = \max_{X \in P} \prod_{i=1}^{n} (X_i - d_i)$$

In the general case, however, the disagreement vector is not known in advance, and must be worked out by the players as part of the bargaining problem. But the system of equations given above would then have 3n unknowns, and so must be augmented if a solution is to be found. The natural place to look for further constraints is to the bargains reached among the syndicates, since the unique property of n-person cooperative games is that groups of players can form freely and act to protect common interests. The overriding problem, however, is to introduce constraining equations which represent rational constraints on the bargaining taking place among the n players. That is, although there may be any number of ways to augment the basic set of equations so as to make the resulting system solvable, the additions must be defensible as considerations any rational player would take into account.

There are two types of constraints Harsanyi introduces. The first constitutes the introduction of equations representing the interdependencies of bargaining among the various syndicates. The second are mathematical conveniences introduced to make the system work. Each will now be discussed.

6.5.3. Type 1 Constraints

6.5.3.1. The dividends

Since Harsanyi defines the dividends guaranteed by the syndicates to the players as being additive, we can unite

player i's final payoff as the sum of the dividends paid him by each of the syndicates of which he is a member:

$$(6.15)\quad u_i = \sum_{\substack{S \subseteq N \\ i \in S}} w_i^S$$

where w_i^S = the dividend guaranteed to player i from syndicate S

Clearly, the sets of w_i^S for each player i become the "solution unknowns", replacing the u_i (the final payoffs in the game) which are now simply defined variables. We must also introduce constraints on these "dividends", however. Harsanyi suggests the following:

> We want to exclude dividend guaranteeing agreements which are in themselves unrealistic or are inconsistent with the other commitments of the participants. Therefore, we shall require that the dividends guaranteed to any player i by a given syndicate S and by all its subsets taken together should not exceed the conflict payoff u_i^S which is the largest payoff that the cooperative effort of the members of S could secure for him in the event of a conflict between S and $\bar{S}$. On the other hand, we also want to exclude dividend-guaranteeing agreements which are inefficient (not Pareto-optimal), and therefore also require that the sum of these dividends should not fall short of this conflict payoff, either.[33]

That is, Harsanyi suggests that the following constraint be placed on the dividends offered by syndicate S and its subsets to player i:

$$(6.16)\quad u_i^S = \sum_{\substack{R \subseteq S \\ i \in R}} w_i^R$$

where u_i^S = the conflict payoff to player i from syndicate S when S is opposed by $\bar{S}$ and both play their conflict strategies

Of course, as it stands, the u_i^S are merely variables defined in terms of the dividends. The crux of the matter is, however, that since these conflict payoffs represent outcomes in the original game given the choice of certain "conflict" strategies by S and $\bar{S}$, an independent derivation of their value can be made on the basis of the optimal choice of threat strategy by each coalition. These equations, then, will eventually serve to constrain the dividends themselves.[34]

6.5.3.2. Dividend proportionality

In the case of a known disagreement vector $d = (d_1, \ldots d_n)$, the "dividends" guaranteed to two players by the all-member syndicate (all others being irrelevant due to the disagreement vector's being known) are in the ratio:

$$(6.17) \quad \frac{u_i - d_i}{u_j \; d_j} = \frac{a_i}{a_j} = \frac{H_i(\bar{u})}{H_j(\bar{u})}$$

Harsanyi wants to show that no matter what syndicates players i and j are both members of, the ratio of their dividends from those syndicates will be this same value, i.e. $\frac{a_i}{a_j}$

This is one-half of Harsanyi's attempt to reduce the n-person interaction to a series of two-person bargaining games. The other will arise in the next section.

Consider the two-person game Γ_{ij} in which players i and j are to decide upon the distribution of dividends between them for all syndicates of which they are both members. Assume that all other dividends have already been decided; that is, that all final payoffs for all other players are already awarded, and that all dividends to players i and j from syndicates not involving both have also been decided on.

Harsanyi contends that the disagreement vector for this two-person subgame is given by:[35]

$$(6.18)\quad t = (t_i, t_j) = \Big(\sum_{\substack{S\subseteq N \\ i\in S \\ j\notin S}} w_i^S,\ \sum_{\substack{S\subseteq N \\ i\notin S \\ j\in S}} w_j^S\Big)$$

He also wants to contend that since all other agreements can be taken as resolved, that the final payoffs from Γ_{ij} to players i and j are those players' final payoffs in the original game Γ.

With these assumptions, the Nash model provides a unique solution to Γ_{ij}, and the ratio of payoffs is, as desired, $\frac{a_i}{a_j}$.

But Harsanyi wants to show that the dividends to player i and j are in this ratio in each of the syndicates S containing them both, not just in the game Γ_{ij}. To do this

he shows that the game involving the bargaining over the ratio of dividends in some particular syndicate S can be regarded as a smaller subgame of Γ_{ij}, and the desired result follows from a general theorem concerning how rational players play such subgames. Intuitively we may think of this theorem as saying that if Γ_{ij} is to be played in two stages with the players bargaining first over the payoffs for some S of which both are members and then those payoffs are used as the disagreement payoffs for Γ_{ij}, then rational bargaining will proceed in such a way as to leave the players' bargaining positions in Γ_{ij} unaltered. What this means is that the payoff ratio in each "subgame" will be the same as it is for the game Γ_{ij}, and Harsanyi's desired conclusion follows immediately.[36]

6.5.3.3. Mutually optimal conflict strategies

With the two previous sets of constraints, the equations defining the solution to the n-person game can be written:

$$(6.19)\quad H(\bar{u}) = 0$$

$$(6.20)\quad a_i = H_i(\bar{u}) \text{ for all } i \in N$$

$$(6.21)\quad a_i w_i^S = a_j w_j^S \text{ for all } i, j \in N$$

$$(6.22)\quad u_i^S = \sum_{\substack{i \in R \\ R \subseteq S}} w_i^R \text{ for all } i \in N$$

This set is solvable for a solution vector $\bar{u} = (u_1^N, u_2^N, \ldots, u_n^N)$ if we have independent definitions of the

"conflict payoffs" u_i^S. Harsanyi arrives at these by considering the choice of mutually optimal threat strategies by each pair of opposing coalitions. This maneuver introduces the second half of the reduction of the n-person game to two-person subgames.[37]

Harsanyi treats the choice of mutually optimal threat strategies as two-person "threat" games ($\tilde{\Gamma}_{ij}^{S,\bar{S}}$) involving any one player from each of the opposing syndicates, say player i from S and player j from $\bar{S}$. Each player is to choose a strategy ($\theta^S, \theta^{\bar{S}}$) from among his syndicate's set of joint strategy options ($\Theta^S, \Theta^{\bar{S}}$) so as to maximize <u>his own final payoff</u> (u_i, u_j). The important thing for Harsanyi to prove is that this threat game is treatable as a two-person zero-sum game the solution to which is independent of the choice of player i and player j as representatives of their respective syndicates. If so, then the conflict payoffs u_i^S are defined, and the system of equations can be solved for the desired solution vector $\bar{u}$.

Although the details of Harsanyi's derivation are quite complex and need not be repeated for our purposes, he was able to prove this important result, and so to complete the reduction of the n-person game to two-person subgames. The only problem is that the resulting system of equations is, in general, non-analytic, and must be solved iteratively.[38] The following representation of his proposed solution to the threat subgame $\tilde{\Gamma}_{ij}^{S,\bar{S}}$ shows this non-analytic property clearly:

$$(6.23)\quad \sum_{i\in S} a_i u_i^S - \sum_{j\in \bar{S}} a_j u_j^{\bar{S}} =$$

$$\max_{\theta^S \in \Theta^S} \min_{\theta^{\bar{S}} \in \Theta^{\bar{S}}} \left[\sum_{i\in S} a_i X_i(\theta^S,\theta^{\bar{S}}) - \sum_{j\in \bar{S}} a_j X_j(\theta^S,\theta^{\bar{S}}) \right]$$

subject to:

$$(6.24)\quad \begin{aligned} a_i w_i^S &= a_k w_k^S \qquad i,k\in S \\ a_j w_j^{\bar{S}} &= a_m w_m^{\bar{S}} \qquad j,m\in \bar{S} \end{aligned}$$

where $X(\sigma) = (X_1(\sigma), X_2(\sigma), \ldots, X_n(\sigma))$ = a payoff vector for Γ $(\in P)$

$\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ = a strategy vector for Γ

Equations (6.19) through (6.24), then, represent the system Harsanyi proposes as offering a solution concept for the n-person cooperative game. However, in deriving his results and in proving that a mathematical solution to the system always exists, Harsanyi was forced to make a number of simplifying mathematical assumptions. Before proceeding to Harsanyi's definition of social power, I want to outline briefly what some of these assumptions were, as they will arise in later discussion.

6.5.4. Type 2 Constraints

6.5.4.1. Introduction

Any considerations introduced in the development of the above equation system or in the demonstration that the system has certain properties (eg. that a solution exists), are important either because they reveal assumptions Harsanyi is making about the nature of rational behavior, or because they restrict the generality of his proposed solution concept. This is true even of purely "formal" constraints along the lines of "let us assume that first derivatives exist".[39]

Harsanyi makes three assumptions of this type worthy of note. He assumes that syndicates can declare negative dividends, that if the derivatives H_i do not exist we may without distortion approximate that game by ones in which they do, and that a system of equations which is non-analytic can still define a legitimate solution concept. I will now briefly explain the importance of each of these assumptions.

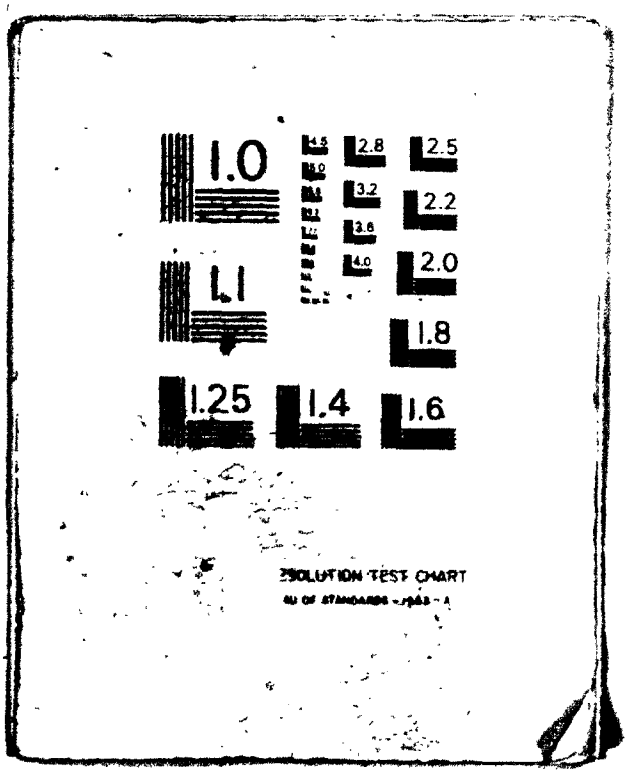
6.5.4.2. Negative dividends

Harsanyi realizes that his approach does not rule out the possibility that:

$$\sum_{\substack{i\in S \\ R\subset S}} w_i^R > u_i^S \qquad (6.25)$$

in which case the sectional syndicates R will have, as it

3



were, "overspent their budget" for player i. To resolve this apparent conflict with the constraint (6.22), Harsanyi is committed to allowing syndicate S to declare a "negative dividend" to player i. The question is whether such negative dividends, though mathematically unexceptional, would be tolerated by rational players. Moreover, could we expect the dividend proportionality rule to apply in such cases?

Although Harsanyi does not offer an answer to the first question, his response would likely be that the players in $\bar{S}$ would <u>force</u> the members of S to suffer the negative dividends. The real constraints for the members of S are the low conflict payoffs u_i^S over which they have no control; the constraints are not really the negative dividends themselves which we might naively think the players could choose to reject. If the members of S chose not to form a syndicate, the members of $\bar{S}$ would likely be able to force the "conflict" payoffs even lower, since the members of $\bar{S}$ would not have the opportunity to act in concert. Since they do not want that, the members of S must form a coalition and must accept negative dividends.

As to whether the negative dividends should obey the dividend proportionality rule, Harsanyi suggests the following rationale. Given that the sum of the players' dividends has to be negative and that it is left up to the players to decide how to distribute this loss among themselves, rational players will bargain their way to a distribution which leaves

$\frac{t^* - r^*}{X^*}$ is the relative strength of B's power over A. (That is, his power to get A to tolerate his performing not-Q, which A will now have to do with the probability $(1-p_2)$.) It represents the opportunity cost (relative to X^*) to A of a conflict with B.[50]

The difference between these two relative strengths of power, then, is the net strength of A's power over B. Harsanyi notes that this measure differs from Dahl's by a simple factor of 2.[51]

6.6.2. Harsanyi's n-Person Model

The important features of Harsanyi's treatment of n-person reciprocal power situations are those which enable him to reduce the interaction to a form amenable to solution as an n-person cooperative game. As they arise in the general case, these features are the following:

1) The n agents have (n+m) pure joint policy alternatives X_j.[52]

2) Each agent i assigns an expected utility u_{ij} to each policy X_j. These entail a greater or lesser conflict of interest over what joint policy is to be adopted.

3) In the event of an agreement being reached on what joint policy is to be enacted (call it $P_{opt} = (p_1, p_2, \ldots p_{n+m})$ where p_j = the probability of the n players jointly enacting policy X_j) the agents each pay each other a "reward" given by the overall reward strategy ρ. (Each agent i has a "net" reward r_i given by the net expected utility of having to pay his promised rewards to each of the other players and the expected utilities of the rewards he receives from each of the other players.)

Even if we do not question whether these equations can be derived in the manner suggested, we should ask whether the result is at all meaningful to the participants in the original games.

However, since Harsanyi offers no rationale for this introduction of a sequence of games, he seems to be suggesting that he regards it as a "merely formal" maneuver, of no more concern to rational players than a theorem in arithmetic. They will undertake this form of reasoning because it is "rational" in the sense of logic, not in the sense of utility maximization. Therefore, it requires, for him, no special justification.[41]

6.5.4.4. Non-analytic systems of equations

A non-analytic system of equations cannot be solved "directly" for the unknowns. That is, there is no analytic means of determining what the solutions to the system are. Moreover, in general there is no way of knowing how many solutions there are that satisfy all the equations. Finally, since non-analytic systems must be solved iteratively, they must be provided with an "initial guess", and the "solution" towards which the iterations converge is (in general) a function of that initial guess.[42]

In his initial paper on cooperative games, Harsanyi suggested that although in general his model will generate a number of "solutions" to the n-person game, rational players will realize that bargaining must again be undertaken

to determine which of these is to be accepted as the final "stable" solution.[43]

Although he does not really tell us why, by his second paper Harsanyi moved away from this position and seemed willing to accept the non-uniqueness of his solution concept.[44] I want to suggest what I think might have been a reason why this shift occurred.

One of the interesting properties of non-analytic systems of equations is that you can never tell when you have discovered all the existing solutions. This being so, Harsanyi's second round of bargaining could never have taken place. Rational players would in general continue to look for more solutions to the original game rather than turn to the second round of bargaining. After all, in the purity of game theory time is no constraint, and a given player may never be able to tell whether a yet-to-be-discovered solution would improve his bargaining position. Where there is no rush to find a solution, the players will continue their hunt for alternatives. Although this argument seems to suggest that we should restrict ourselves to the first round of bargaining in order to satisfy descriptive criteria of adequacy, I think it says more. The second round of bargaining is <u>in principle</u> unattainable for the general n-person game. It is not merely a matter of the finite reasoning capabilities of the players. So when Harsanyi suggests that "For empirical applications...it now seems to me the various

solutions...satisfying [the system of equations] are much more interesting concepts than the "stable solution", he is either missing the main point, or is misleading his readers as to the real motive for the return to the set of solutions.

6.6. Harsanyi's Theory of Social Power

6.6.1. Introduction: Harsanyi's Two-Person Model

As outlined in section 6.1 Harsanyi's theory of social power builds on an analysis of the opportunity costs involved in an influence attempt. In a two-person interaction his model can be described very simply.

Consider an agent B who is about to enact some policy Q with probability p_1, and therefore the policy not-Q with the probability $(1-p_1)$. Call this mixed strategy s_1. Let B's expected utility in this case be given by $u^o(s_1)$ where:[45]

$$(6.26) \quad u^o(s_1) = u_1 - p_1 X$$

and where u_1 and X are parameters defining B's utility function and its sensitivity to the probability of his performing policy Q.

Now agent A comes along and threatens to punish B if he performs his previously chosen strategy s_1 but to reward him if he increases the probability of his performing policy Q. That is, he will reward him if he performs the mixed strategy $s_2=(p_2,(1-p_2))$ for some yet to be agreed upon value

$p_1 < p_2 \leq 1$.

Agent B now has two expected utilities to consider:

(6.27) $u(s_1) = u_1 - p_1X - t$

(6.28) $u(s_2) = u_1 - p_2X + r$

where t = the expected utility of the threatened punishment

r = the expected utility of the promised reward.

Of course, we would expect agent A to have good reason for making these promises to agent B. With no loss of generality we can assume that A wants B to perform policy Q, or at least to increase the probability of his performing Q. That is, that the more likely it is that B enacts Q, the higher is A's expected utility. But, of course, the rewards and punishments are <u>also</u> costly to A. That is, even though a higher value for p_2 is in itself desirable to A, the offering of a reward for compliance reduces his net expected utility even if B should comply, while the threatened punishment would reduce his expected utility even further in the event B refuses to change. Agent A, then, <u>also</u> has two expected utilities to consider:

(6.29) $u^*(s_1) = u_1^* + p_1X^* - t^*$

(6.30) $u^*(s_2) = u_1^* + p_2X^* - r^*$

where u_1^* and X^* are parameters defining A's utility function and its sensitivity to the probability of B's performing policy Q

and where t^* = the expected utility of punishing B

r^* = the expected utility of rewarding B

Now, if both players were rational, they would agree upon some value for p_2 intermediate between the probabilities corresponding to what Harsanyi calls the two players "concession limits". That is, they would agree to a value for p_2 intermediate between, at the high end, the probability which would make further concession by B irrational, and, at the low end, that which would make further concession by A irrational. (The former is given by that value for p_2 at which $u(s_2)=u(s_1)$, while the latter is given by that value for p_2 at which $u^*(s_2)=u^*(s_1)$.)

Given the parameters of the game and the players' utility functions, the outcome of rational bargaining has the very simple form:[46]

$$(6.31) \quad p_2 = p_1 + \frac{r+t}{2X} + \frac{r^*-t^*}{2X^*}$$

Substituting this value for p_2 into equations (6.28) and (6.30) gives the expected utilities for A and B of this rational resolution of the power struggle:[47]

$$(6.32) \quad u(s_2) = u_o + \frac{r-t}{2} - \frac{X}{2X^*}(r^*-t^*) - p_1X$$

$$(6.33)\quad u^*(s_2) = u_o^* - \frac{r^*+t^*}{2} + \frac{X^*}{2X}(r+t) + p_1X^*$$

Of course, A is also faced with the decision as to what rewards and punishments he should promise B to ensure this optimal outcome. Equation (6.33) shows us that his expected utility is maximized if the reward is chosen so as to maximize the expression:[48]

$$(6.34)\quad \Delta r = \frac{r}{X} - \frac{r^*}{X^*}$$

while it is also maximized if the threat is chosen so as to maximize the expression:

$$(6.35)\quad \Delta t = \frac{t}{X} - \frac{t^*}{X^*}$$

If A has perfect information about B's utility function, these equations can easily be solved for the maximizing reward and punishment.

Now, in a situation like this, Dahl's measure of the "amount" of A's power over B with respect to policy Q is given by the change in the probability of B's performing Q. With the above value for p_2, this measure reduces to:[49]

$$(6.36)\quad \Delta p = p_2 - p_1 = \frac{1}{2}\left(\frac{r+t}{X} - \frac{t^*-r^*}{X^*}\right)$$

Harsanyi suggests that in this expression $\frac{r+t}{X}$ represents the relative strength of A's power over B; that is, it represents the total opportunity costs (relative to X) to B of choosing non-compliance instead of compliance. Likewise, the factor

$\frac{t^* - r^*}{X^*}$ is the relative strength of B's power over A. (That is, his power to get A to tolerate his performing not-Q, which A will now have to do with the probability $(1-p_2)$.) It represents the opportunity cost (relative to X^*) to A of a conflict with B.[50]

The difference between these two relative strengths of power, then, is the net strength of A's power over B. Harsanyi notes that this measure differs from Dahl's by a simple factor of 2.[51]

6.6.2. Harsanyi's n-Person Model

The important features of Harsanyi's treatment of n-person reciprocal power situations are those which enable him to reduce the interaction to a form amenable to solution as an n-person cooperative game. As they arise in the general case, these features are the following:

1) The n agents have (n+m) pure joint policy alternatives X_j.[52]

2) Each agent i assigns an expected utility u_{ij} to each policy X_j. These entail a greater or lesser conflict of interest over what joint policy is to be adopted.

3) In the event of an agreement being reached on what joint policy is to be enacted (call it $P_{opt} = (p_1, p_2, \ldots p_{n+m})$ where p_j = the probability of the n players jointly enacting policy X_j) the agents each pay each other a "reward" given by the overall reward strategy ρ. (Each agent i has a "net" reward r_i given by the net expected utility of having to pay his promised rewards to each of the other players and the expected utilities of the rewards he receives from each of the other players.)

4) In the event of a conflict between two coalitions S and $\bar{S}$, S gets to choose the joint strategy with probability Π_S and $\bar{S}$ gets to choose with probability $(1-\Pi_S)$. If S wants to enact each alternative X_j with probability q_j^S, and $\bar{S}$ with probability $q_j^{\bar{S}}$, then the "joint" conflict strategy has:

(6.37) $$p_j^S = p_j^{\bar{S}} = q_j^S \Pi_S + (1-\Pi_S)\, q_j^{\bar{S}}$$

5) Retaliatory strategies are announced by each coalition S. In the event of a conflict between S and $\bar{S}$ these punish the members of the opposing coalition. Player i's net loss due to this conflict is given by t_i^S.

Under these assumptions, Harsanyi's solution concept for n-person cooperative games provides equations which can be solved iteratively for the rational joint policy P_{opt}, for the optimum net rewards $(r_1, r_2, \ldots, r_n)$, for the net conflict punishment vectors $(t_1^S, \ldots t_n^S)$ for each coalition, and for the joint policies in the event of conflict.

Given these results, Harsanyi proposes to define the strength of an individual i's power over the joint policy as that value $\overline{\overline{p_i}}$ for which

(6.38) $$\overline{\overline{p_i}} \max_{j\in(N\cup M)} (u_{ij}) + (1-\overline{\overline{p_i}}) \min_{j\in(N\cup M)} (u_{ij}) = u_i(P_{opt}, \rho)$$

where u_i = player i's net expected utility of the joint policy P_{opt} and reward strategy ρ. That is, it is the probability at which the expected utility of a mix between the policy i most favors and that which he least favors equals the expected utility of the jointly agreed upon strategy P_{opt}.

6.7. Critique of Harsanyi's Theory of Cooperative Games

6.7.1. Introduction

In Chapter IV it was suggested that in his theory of non-cooperative games Harsanyi was committed to satisfying three criteria of adequacy: Formal Existence and Uniqueness, Normative Interpretability, and Psychological Interpretability. With the exception of his rejection of the uniqueness requirement, these same criteria apply to his proposals in cooperative theory. Of particular importance, of course, is the satisfaction of the "positive" interpretation, since we are here considering his theory's ability to support an explanatory notion -- social power.

As in that earlier analysis, there is little doubt that the system of equations Harsanyi proposes as his solution in cooperative theory has the formal properties he claims for it. He set out to generate equations representing certain equilibrium conditions and in that he succeeded. If his theory is lacking anything it is surely not formal elegance.

However, I want to argue again that Harsanyi's proposals fail the other two tests they are committed to passing. In short, I will argue that some of the moves Harsanyi makes in developing his system of equations are not sufficiently well grounded for his proposal to stand as a normative theory, and that the overall system he proposes as a solution to conflict between rational agents lacks a legitimate

psychological model.

6.7.2. Failure of the Normative Model

A normative theory of the sort Harsanyi proposes in his theory of cooperative games can fail in two ways. Type 1 failure is of the sort where some of the constraints imposed in the derivation of the normative concept (in this case the concept of a "rational solution") either conflict with our intuitions or are simply unjustified as they stand. Type 2 failure requires that the set of constraints do not in concert satisfy some important pre-systematic intuitions regarding the concept at issue.[53]

In general, type 2 errors are less devastating. If we have a number of alternative theories explicating a normative concept, each one rigorously developed in accordance with alternative sets of intutions, philosophical insight would still be possible. For example, there would still be the possibility that we would judge one of the alternatives to be "better" than the others and thereby gain insight into the relative importance of our intuitions. Moreover, in general it is the case that the bringing together of intuitions in the development of a rigorous normative concept is itself a revealing exercise, even if we are not totally happy with the result.[54]

On the other hand, type 1 errors involve the joining of justifiable assumptions with other formal constraints in conflict with our intuitions or at best with non-sequitors.

The result is clearly not likely to lead to much normative insight beyond that captured by the previously justified assumptions.

I think Harsanyi's proposals fail in both ways, but by far the more important failures are of type 1.[55] Specifically, he imports what I will argue are weakly motivated criteria as constraints on the rational resolution of n-person conflict and offers the result as a normative theory of rational interaction.

We have seen that n-person cooperative games have unique solutions whenever the rules of the game stipulate a unique "disagreement" vector. To derive that solution, moreover, we need not consider each player's bargaining strength, and coalitions do not even arise. From this it follows that the point to the introduction of coalitions (or syndicates) in the general case is, in effect, to come up with constraints which do the job of the disagreement vector. That is, we must replace a normatively neutral boundary condition with internally generated constraints having a normative justification.

As noted above, in pursuing this problem in n-person theory, Harsanyi followed Nash's approach to two-person theory, and assumed that what rational players pay attention to in such cases is "threat potential"; that is, how much "harm" an agent can do to his opponents while not "harming" himself. Now, there are problems with this concept of "threat

potential" even in the normative use of two-person game theory. The main argument against the notion of "threat" is that their incorporation into the theory appears to be in conflict with the assumed rationality of the players.

Why should rational players of a cooperative game choose to agree upon a solution as a function of threat strategies each player knows the other would never play? To be sure, the players both know these particular strategy options are available and if played would have the advertized effects, and both might be assumed to know Nash's theory whereby these "conflict payoffs" would support a certain bargaining "solution". But what is not clear is why rational players should be concerned with either of these facts.[56]

In the traditional theory, von-Neumann and Morgenstern suggest that if the two players are allowed to agree on a joint strategy giving each a higher payoff than each can do on his own, there are no further rational constraints we can impose.[57] That is, if the players are rational they will necessarily agree to a joint strategy giving some payoff pair in the negotiation set of the game.

Of course, threat strategies have been introduced to make the solution unique, but rational players know this. There is, quite simply, no substance to each player "uttering his threat". The only way threats could become relevant is if they are credible, that is if rational players were

sometimes forced to play them. By assumption they are not, and therefore "threat potential" cannot possibly be relevant to the cooperative resolution of conflict between rational players.

The only way around this argument is to give rational players the power to actually play their threat strategies. Thus we must either develop a dynamic bargaining model where the status quo point (perhaps given by mutually optimal threat strategies) has "real significance" to the players as the perhaps "permanent", but at least "present", state of affairs, or we must weaken the notion of rationality sufficiently to allow the players to rationally expect non-utility maximizing behavior from each other. Either option introduces an element of uncertainty regarding the interaction which game theory has abstracted away. That is, both options run counter to what Harsanyi has introduced as the "rationality postulates" characterizing the Nash-Zeuthan approach to two-person cooperative games.[58] (These were proposed for the case of given disagreement vector.) In particular, what would have to be altered are his symmetry postulate, his restriction of variables postulate, and his mutually expected rationality postulate -- in short, all those postulates intended to explicate what expectations rational players might formulate regarding their opponent's likely behavior based on the assumed-known context of interaction.[59]

Note that this argument against the relevance of strategic reasoning is not simply that the assumptions made in game theory are unrealistic, or that by weakening them we may discover a more interesting model of human behavior, but rather that when applied to the two-person case the rationality postulates lying behind Harsanyi's model make any progress beyond the von-Neumann-Morgenstern solution impossible. The players' uncertainty about which element in the negotiation set will be agreed to is not sufficient to make their threats credible, and so in game theory, although surely not in rational interaction, threats are strictly speaking ad hoc.

However these alternative models and the altering of the rationality postulates might be pursued in the two-person case, Harsanyi confronts even greater problems with his n-person theory. As has been noted, coalitions are brought into the picture to help compute each player's bargaining strength, since Harsanyi feels that that attribute is a function of what the agent can do to the opposition acting both by himself and in concert with others. Aside from a generalization of the argument just outlined to the effect that in game theory sectional coalitions will never form anyway and so are irrelevant, the main inadequacy of Harsanyi's normative proposals for the n-person case arises from the details of his treatment of rational coalition formation and its relevance to an individual's bargaining strength.[60] Although

consideration of these details does not lead to as sweeping a criticism of the program as the just-outlined general critique of game theory's ability to deal with strategic reasoning, they nevertheless support that argument by revealing how difficult it is to justify the step-by-step importation of additional constraints on rational behavior, even if threats <u>are</u> rationally credible.

Consider first the disagreement vector for the game Γ_{ij} in which players i and j have to decide on the ratio of their payoffs in all syndicates of which both are members. This vector was given by equation (6.18):

$$t = \left(\sum_{\substack{S \subseteq N \\ i \in S \\ j \notin S}} w_i^S, \quad \sum_{\substack{S \subseteq N \\ j \in S \\ i \notin S}} w_j^S \right)$$

Harsanyi's argument in support of these values was that since we are seeking equilibrium conditions we may consider all other agreements as having been concluded, and so in effect players i and j have <u>already been awarded</u> each of these dividends.

This defence is weak for two reasons. First, strictly speaking Harsanyi's own threat-based treatment of the disagreement vector does not make it equal to "what the players have in hand". The game Γ_{ij} has a large number of alternative proposals open to both agents. These, in turn, are a function of what their strategy options are in the original game Γ, especially what affects these strategies have on

each other's payoffs. If it arises, real disagreement as to how to split their payoffs will not simply result in the two players keeping what they have from the other syndicates and walking away clear from Γ_{ij}. Therefore, it seems ridiculous to assume that these are the conflict payoffs.[61]

Secondly, by the assumed structure of the syndicates, if players i and j cannot agree to get along, then no syndicate of which they are members has any right to declare dividends to anyone. The declaration of dividends presupposes that the members of the syndicate can play <u>any</u> joint strategy to defend their common interest. If players i and j are at odds, however, the joint strategy options for the whole syndicate are restricted, and so no guarantees should have been made.

By assuming as a given that the other players all have come to agreements on how to proportion their payoffs, Harsanyi has assumed enough to show that players i and j should apportion theirs according to some fixed rule. But the only grounds Harsanyi has for <u>making</u> that assumption is his claim that he is only after "equilibrium conditions". But equilibrium points are irrelevant in a normative theory unless they can be shown to have some rationale. Harsanyi's inability to justify the solution to any <u>particular</u> Γ_{ij} without assuming that other similar games have already been justified betrays this lack of a rationale.[62]

Connected with this problem is Harsanyi's defence of

"negative dividends". Following up on his treatment of Γ_{ij}, what Harsanyi is suggesting is that a disagreement vector for a syndicate may legitimately award to all members of the syndicate a greater payoff than they could rationally expect from conflict with the counter-syndicate. This is evident in equation (6.16):

$$u_i^S = \sum_{\substack{R \subseteq S \\ i \in R}} w_i^R$$

where it is clear that if w_i^S is negative, then:

$$(6.39) \quad u_i^S < \sum_{\substack{R \subset S \\ i \in R}} w_i^R$$

But in keeping with his definition of disagreement payoffs in Γ_{ij}, Harsanyi treats the expression on the right hand side of this equation as in effect a form of disagreement payoff for player i in the sectional syndicate confrontation between S and $\bar{S}$. It is "what he has in his pocket" at the time of the confrontation, and represents in payoff terms the power he has accumulated from the subsets of S. But if this expression <u>is</u> a legitimate disagreement payoff, and if the confrontation between S and $\bar{S}$ is in any sense supposed to be "real", then there is no justification for equation (6.39) whereby player i takes a conflict payoff from S which is less than his disagreement payoff.[63]

Of course, Harsanyi may respond that the right hand

side expression represents the power of i in the syndicate S, not his disagreement payoff should S conflict with $\bar{S}$. Consequently, if the syndicate as a whole has negative power (as indicated by the low value of u_i^S) then i's membership in S cannot be to his advantage -- that is, it cannot lead to an increase in his power in the supersets of S. Hence his dividend from S must be negative.

This response has one major flaw. If the right hand side expression does denote "player i's power in S" and not his power "as a member of S", then there is even less justification for negative dividends obeying the dividend proportionality rule. If a syndicate S happens to be weak, and therefore its members are forced to give up dividends already in hand, then its stronger members will force its weaker ones to take most of the punishment. That is, those members with higher "power in S" will suffer less from conflict with $\bar{S}$ than will those members whose "power in S" is low. A weak syndicate may be unable to mount much of a threat against the counter-coalition, but its strong members can nevertheless mount a considerable threat against its weaker ones, and therefore can be expected to suffer less from a disastrous conflict, and not more as the dividend proportionality rule requires.

Another problem in detail concerns Harsanyi's unabashed extension of the prospect space P (and the corresponding strategy space) of the game to include all payoff vectors

weakly dominated by members of P (and some corresponding strategies).[64]

Now, Harsanyi suggests that this assumption amounts to saying that any player can voluntarily reduce his payoff any amount. But since there would appear to be no possible rationale which would lead rational players of the original game to want to do this, i.e. to deal in the extended spaces, this extension would seem to be blatantly ad hoc.

On the other hand, if the effects of the assumption were in principle eliminable and it served only to simplify the mathematics, only a purist would object. Unfortunately, the assumption is central to two of Harsanyi's proofs. It arises in his showing that $\tilde{r}_{ij}^{S,\bar{S}}$ has a solution, and in his proof that a solution exists to the overall system of equations.

Whether the assumption of an extended prospect space is essential to these results is hard to tell, but until we see that it is not, the results themselves appear as ad hoc as the assumption.

Finally, I want to consider the equilibrium approach as a method for generating normative concepts. What Harsanyi has tried to do is to generate a set of equations defining the constraints imposed by rational agreements on each other, when the agreements are reached among subsets of a group of players each of whose eventual choice of strategy will (in general) affect everyone's payoff. Now, when a disagreement

vector is given, and if we assume it to have credibility, then it serves legitimately as a boundary condition on the reasoning of rational players. Under the same assumption, the security levels of the players might also be considered as a legitimate boundary condition of sorts. In the general case, however, there is no boundary condition, and so Harsanyi has tried to generate one by considering "equilibrium constraints."

Now, since equilibrium is in itself neither desirable nor undesirable to rational players, Harsanyi has had to rationalize each step in the generation of these constraints. But even if we grant him success in this regard, because of his overriding concern with equilibrium each such defence will be of the form: "Given that the other agreements have been concluded rationally, rationality compels the players to conclude this agreement in the following manner...". That is, no agreement would be by itself and unconditionally rational, since the "boundary conditions" would be formed by the expectations of the players regarding the other agreements, and none of these would be unconditionally rational. Since the players expect nothing but rationality from each other, they could not have even begun to generate the first agreement. The effect is that even if Harsanyi's system does give equilibrium conditions among the agreements, and even if each agreement is rational "given the others", the system still lacks sufficient support to stand as an unconditional normative theory under the usual assumption of

mutual expectation of rationality.[65]

Additional support for the equilibrium model may come from a number of sources: the boundary conditions mentioned above, an assumed sequence of coalition formation; or a weakening of the assumptions lying behind an agent's rational expectations regarding his opponent's behavior. Each might be of some help in some cases. In general, however, something has to augment mere equilibrium for any system of equations representing those equilibrium constraints to function as a definition of a normative concept like "rational solution".

6.7.3. Failure of the Psychological Model

The psychological interpretability criterion requires that a viable solution must supply a "plausible psychological model for the actual bargaining process".[66] For Harsanyi it is not enough that a proposed solution satisfy certain axiomatic expressions of rational constraints, since it must also be representable as the culmination of some rational reasoning process in order to function in positive contexts.

For two-person cooperative games this "plausible psychological model" supporting the Nash axiomatic approach is supplied by Zeuthan's solution to the bargaining problem.[67] In effect, that solution models the bargaining process as a series of concession made by the opposing players, with the decision as to which player ought to "give in" at each

stage being based on "objective" estimates of how likely it is that the opponent is going to give in. The series of concessions made in this way advances the players from the status quo point to the same final agreement given by the Nash axioms.

The issue here is whether a psychological model is possible for Harsanyi's n-person generalization of the Nash model. I want to suggest that there are three reasons why it is unlikely that a general psychological model providing the same solution can be found.

First, no real psychological process can model Harsanyi's n-person proposal since in general there is no analytic definition of a starting point for the bargaining process. This was not a problem in the two-person case since mutually optimal threats are defined analytically. Psychologically, of course, agents cannot help but have a "starting point" in an interaction since they always have a history and always have expectations. To be sure, the expectations may be revised as they interact, but there would always be some initial beliefs and a starting status quo point, and as such they would remain unaltered.

More generally, because the solution equations are non-analytic, it is unlikely that there is an equivalent but analytic set of equations defining the same set of solutions.[68] This being so, there can be no analytic procedure a rational agent can follow to arrive at a solution to the

bargaining problem. Hence, if he is to "solve" the problem with an answer among those Harsanyi's system allows, the rational agent must simply apply iterative techniques with the sole purpose of solving that same system of equations. But this is tantamount to his assuming that the solution offered by Harsanyi is correct and that the rational thing to do is simply to solve his system of equations.

Finally, there is the problem of the non-uniqueness of Harsanyi's solution concept. One of the strange things about conscious human thought processes is that we can only think one thing at a time. Thus, if there is a psychological model for n-person bargaining, not only must it have a unique starting point, but if it is to have any end point at all, that must be unique also.

Now it might be suggested that a unique end point might be of the form "any of the following joint actions would be rational". But as Harsanyi points out, none of the actions in such a set would be agreed to by all the players, since none would be "stable", and another round of bargaining would have to follow. But more importantly, as I have suggested, the iterative requirement for solution of the system of equations leaves the number of solutions unknowable, and so such an end point could never really be reached by <u>any</u> process.

Thus, not only has Harsanyi not provided us with a plausible psychological model of the bargaining process,

but in general such a model would not appear to be forthcoming. Therefore, to make the theory empirically relevant some assumption must be weakened. Harsanyi suggests that it is "more realistic to assume that the players will usually know the prospect space only in the immediate vicinity of their actual position".[69] While this may be the case, the conclusion we must draw is that some such move is required for more than the sake of "realism". It is demanded if there is to be the possibility of developing any psychological model of the process of bargaining between rational players.

6.8. Critique of Harsanyi's Theory of Social Power

I want now to consider three general arguments against Harsanyi's theory of social power. The first amounts simply to the claim that his treatment of social power is unacceptable unless his theory of cooperative games can overcome the objections raised in the previous sections. This much follows from the direct link between the two.

The second general argument concerns three of the features of his reduction of the n-person policy interaction to the form of a cooperative game.

1) As it stands, his solution provides figures for each player i's net rational reward r_i. This represents what would amount to the utility that player i assigns to the bundle of commodities, promises, and transfers of all sorts, given to and received from the other players of the game as part of the reward strategy ρ. What is <u>not</u> given,

and what must be given if the solution is to have any application, is figures representing what the utility of each reward offer should be. But as Harsanyi himself notes, this is in general impossible because the utility of any player's bundle need bear no particular relation to the utilities of the individual elements. Therefore, his theory offers virtually no advice to each agent as to what rewards to offer the others in order to get them to comply, while if the agents choose their own, there is no guarantee that their power will even come close to the "theoretical optimum".[70]

2) This same criticism applies against the determination of the optimum threats t_i^S.

3) In order to reduce a conflict situation between coalitions S and $\bar{S}$ to one where a specific probability distribution across the policy alternatives will result (that is, to meet the requirements of cooperative games) Harsanyi has to make the following assumption:

> We shall assume that, in case of a conflict between coalitions S and $\bar{S}$, coalition S would have the choice among the policy alternatives $X_1, \ldots X_{n+m}$ with probability Π_S while coalition $\bar{S}$ would have the choice among these alternatives with probability $(1-\Pi_S)$. If the choice were made by coalition S, alternative X_j would be selected with probability q_j^S; if the choice were made by coalition $\bar{S}$, alternative X_j would be selected with probability $q_j^{\bar{S}}$.[71]

Harsanyi then proceeds to define the conflict outcome as being determined by the probabilistic mix of what each coalition would choose should it be given the chance. This was given in equation (6.37).

$$p_j^S = p_j^{\bar{S}} = \Pi_S q_j^S + (1-\Pi_S) q_j^{\bar{S}}$$

But what his solution concept gives him is not what each coalition should do, but rather a value for this probabilistic mix of what the competing coalitions would want to do. That is, we are not given conflict strategies at all.

> Now even if the Π_S are given by the rules of the game and the p_j^S are given by Harsanyi's solution, equation (6.37) cannot be solved for either the q_j^S or the $q_j^{\bar{S}}$. But if we regard the Π_S as themselves variables, then should conflict arise between S and $\bar{S}$ Harsanyi's theory of power has not told us either what policy mix each coalition should want, nor which coalition, if either, gets to enact its favored policy.

The third argument I have against Harsanyi's theory cuts equally well against any other proposal which assumes that power is exercised relative to some fixed, prior choice context.[72] I will register my objection as it applies to the two-person case although the argument applies equally to the n-person model.

Harsanyi assumes that the situation in which power is being exercised can be modelled with the agent being influenced (agent B) having already reached a rational choice of strategy in some decision context. Agent A then steps in and offers him a reward and/or threat.

Now, we might object that many exercises of power take place before the agent actually confronts the prior decision context. But should that be a problem, Harsanyi would contend that we could "theoretically" remove the effects of A's influence to determine what B would have done had A not been present to exercise his power. This is well and good if the direct effects of A's influence are of a certain sort. For example, they might constitute nothing more than augmentation of the consequences of B's alternative strategies. In such a

case, the two choice contexts have the same structure, and only B's utility assignments have been altered.

However, if A has exercised his power in other ways, the "theoretical" construction of the prior choice context for B is not so straightforward. For example, A might have done nothing more than force B into the choice situation itself. Alternatively, he might have provided B with part of his belief function. In such cases, what we should do "theoretically" or otherwise to remove the effects of A's influence and to determine what B "would have done otherwise", is an unexplored issue.[73]

What this final criticism amounts to is an argument to the effect that Harsanyi's theory of social power is at most an incomplete picture. But it also suggests that it fails because it assumes that the question "what would the agent B otherwise have done" can be answered either by fiat with a fixed "prior" probability, or by an examination of the "rewards and threats" A has offered as part of the interaction.

In the next chapter I generalize this criticism and argue that social power is inherently a dynamic notion which no purely statical approach can characterize.

6.9. Conclusion

In this chapter I have summarized and criticized Harsanyi's theory of cooperative games and the theory of social power he constructs on that foundation. Both have been presented to support more general theses regarding the nature of rational interaction.

More specifically, I have attempted to show through a detailed analysis of Harsanyi's proposals that his game-theoretic approach to interdependent rationality is inadequate, both on the strictly normative side, and when it comes to the positive use of game-theoretic notions. That some dual-interpretation approach may be generated to support explanatory notions like social power is still possible. What I have shown, however, is that game theory cannot provide the formal structure for that approach.

FOOTNOTES

1. See Chapter V, footnote 2.

2. See Chapter V, footnotes 7 and 8 and Harsanyi's (1962a) p. 67.

3. Ibid., p. 69.

4. Ibid., p. 69.

5. Ibid., p. 70

6. Ibid., p. 74.

7. Ibid., p. 75.

8. As Rapoport put it in his Two-Person Game Theory (1966), p. 112): "Roughly speaking, Nash's solution favors the player who combines a certain degree of prudence with a certain degree of brinkmanship."

9. The essential Nash papers are his (1950) and his(1953).

10. See, for example, Luce and Raiffa's (1957), pp. 126-127.

11. See his "On the Rationality Postulates Underlying the Theory of Cooperative Games," (1960), and section 6.7.2 below for a discussion of how these postulates must be altered in view of the conflict between assumed rationality and the use of strategic reasoning.

12. This reduction of cooperative games to non-cooperative ones reflects Nash's belief that the latter are more basic. (See Luce and Raiffa (1957), p. 141.) Schelling, for one, is violently opposed to this bias, as is attested to by his (1960).

13. See his (1953) and Luce and Raiffa's discussion in their (1957), pp. 140-143.

14. See Luce and Raiffa (1957), p. 180.

15. Ibid., p. 180-182.

16. An "inessential" game is one in which the weak inequality in equation (6.3) is only satisfied by equalities. In such cases there is no reason for any of the players to form any coalitions. (See Rapoport's N-Person Game Theory (1970), p. 83.)

17. For a further discussion of this concept see Luce and Raiffa (1957), p. 192; Rapoport's (1970), p. 89; and Riker and Ordeschook's An Introduction to Positive Political Theory (1973), p. 134.

18. For example, all constant-sum N-person games have empty cores. See Rapoport's (1970), p. 91.

19. For a further discussion of this concept, see Shapley's (1953) and Rapoport's (1970), p. 106-113.

20. The purpose behind considering the "order of formation" of a coalition was discussed in Chapter V, section 5.3.4.

21. Luce and Raiffa (1957), p. 190.

22. McKinsey (1952), p. 351.

23. Luce and Raiffa (1957), p. 191.

24. Harsanyi (1963), p. 203.

25. Luce and Raiffa (1957), p. 191.

26. Ibid., p. 233.

27. Harsanyi uses this argument in many places. See, for example, his (1956), p. 156, and his (1962a), p. 74.

28. Harsanyi (1963), p. 206.

29. This connection is noted by Rapoport in (1970), Chapter 10.

30. See the Appendix to Harsanyi (1959) and Harsanyi (1963), p. 202-204.

31. Harsanyi (1959), p. 328, and Harsanyi (1963). In what follows I will stick closely to the symbolism used in Harsanyi (1963).

32. Harsanyi (1959), p. 329.

33. Harsanyi (1963), p. 206.

34. See below, section 6.5.3.3.

35. Ibid., p. 207. In fact, Harsanyi says that this is "obvious". For a criticism of this assumption, see below section 6.72.

36. Ibid., p. 208.

37. Ibid., p. 211-214.

38. Harsanyi admits that the constraints make the system "non-recursive" (which I read as non-analytic) in his (1962b), p. 88, but nowhere does he discuss the importance of this fact. For criticisms of his model arising from this property, see below, sections 6.7.2 and 6.7.3.

39. By a "formal" assumption I mean one which Harsanyi seems to feel has no important psychological or normative implications.

40. In section 6.7.2, I argue that rather than accept dividends that accord with this rule, the stronger players should be able to force the weaker ones to accept the lion's share of the punishment. This would (in some way) reverse the ratio for negative dividends.

41. I do not criticize this move in what follows, for it is clear that the consideration of the artificially constructed "sequence of games" is totally ad hoc. In pursuing this line, Harsanyi once more follows Nash (See his (1953), p. 131). But as Luce and Raiffa argue (1957), p. 142, the move is a "completely artificial mathematical escape". This same comment, of course, applies to Harsanyi's tracing procedure, as detailed in Chapter IV.

42. All these are basic properties of non-analytic systems of equations.

43. Harsanyi (1959) p. 347-348. Harsanyi argues there that the "noncontroversial" conflict payoff to player i is the lowest payoff he could get among the various alternative solutions. Why this should be the case is hard to see. In general, this second round of bargaining ought to be undertaken in the same way as the first. But, since such a potential regress would clearly undermine Harsanyi's approach, he must have felt compelled to make this ad hoc assumption at the second stage.

44. Harsanyi (1963), p. 219.

45. This symbolism differs slightly from Harsanyi's in (1962a) for ease of presentation. Moreover, in all of what follows I will assume that A's power over B's behavior is positive, and that at no point do we allow probabilities to be less than 0 or more than 1.

46. See Harsanyi (1962a), p. 77.

47. Ibid., p. 76.

48. Ibid., p. 77.

49. Ibid., p. 77.

50. Ibid., p. 77-78.

51. Ibid., p. 78, Theorem II.

52. This is the general case. In (1962b), Harsanyi treats other cases which I will ignore.

53. This distinction parallels the controversy arising in economics over the "realism" of the assumptions built into a model. In normative models, however, the assumptions are largely what determine the acceptability of the model.

54. A trivial example of this sort is brought about when acceptable axioms lead to an unacceptable result. For instance, Arrow's Impossibility Theorem. See Arrow (1951).

55. The "overall" failure of his concept is primarily a function of the restrictive assumptions it makes regarding the context of interaction. (See Chapter 8).

56. The argument here is simply that since rational players expect rational behavior from their opponents, they would attribute a subjective probability of zero to an opponent's playing a threat strategy. The only justification for the consideration of threats, then, is that they seem to be psychologically relevant. But as I argue below, this psychological relevance builds on a model of rational behavior which is at odds with Harsanyi's. In particular, it weakens the assumption of mutual rationality.

57. See Luce and Raiffa (1957), p. 115-119.

58. Harsanyi (1960).

59. These postulates are (roughly): Symmetry: The same decision rules are adopted by the two players; Restriction of Variables: No variables outside the normal form of representation are relevant; Mutual Rationality: Both players anticipate each other's choice of decision rules to be in keeping with the remainder of the postulates. (See Harsanyi's (1960), p. 184.)

60. This argument runs as follows. The agents can, in general, always improve their payoffs by acting jointly. To act jointly requires that the N-person coalition forms. Therefore, if the players are utility-maximizers, no sectional coalitions could possibly form. Therefore, they are irrelevant.

61. This is just another instance of Harsanyi's assuming, when it suits him to do so, that games have a unique disagreement vector not given by mutually optimal threat strategies. See footnote 43 above.

62. See the last part of this section for a general argument against Harsanyi's pursuit of equilibrium conditions.

63. In effect, a negative dividend under this interpretation of equation (6.39) implies that the conflict payoff is not the payoff resulting from a conflict between S and $\bar{S}$.

64. Harsanyi (1963), p. 204-205.

65. A "conditionally normative theory" as construed here is one which makes a normative prescription of one sort based on normative judgments of another. In this case, the solution Harsanyi proposes is rational if each of the agreements is, but each of those, in turn, is rational only given that the others are (since the way in which the others have been "concluded" forms part of the data base for each player's bargaining).

66. Harsanyi (1956), p. 144.

67. See Zeuthan's (1930) and Harsanyi's discussion in his (1956).

68. Strictly speaking, of course, the set of equations which simply stipulates the alternative solutions is analytic, but it would not count as "equivalent" in my sense since the "rationale" for the axioms of Harsanyi's theory would not have been captured.

69. Harsanyi (1963), p. 219.

70. The point here is that the agent cannot be said even to possess the power unless there is in principle some way for him to determine how to exercise it. In giving only "bundle" values, Harsanyi is leaving the agents with an n-person non-cooperative game to play in order to arrive at their promises. If the analysis of Chapter 4 is any indication, there is reason to doubt that there is any unique solution to this problem, and so there is in principle no way for the players to determine what their reward strategy should be.

71. Harsanyi (1962b), p. 86.

72. Dahl's is the most obvious formulation of this sort. See his (1957).

73. I explore these questions in Chapter VII and Chapter IX.

CHAPTER VII

AN ALTERNATIVE FOCUS FOR SOCIAL POWER

7.1. Overview and Plan of Attack

In the arguments concerning social power in Chapters V and VI and in the earlier chapters on Harsanyi's non-cooperative game theory, I attempted to point out inadequacies in the treatments of social power and rational interaction offered by Harsanyi and Goldman. Although the tone was critical, my objective was constructive. As indicated earlier I have been preparing the way for the presentation of an alternative conceptual framework for handling both the explanatory notion of social power and the related normative issues concerning rational interaction and strategic choice. This alternative is an exchange-theoretic model based on a normative interpretation of the explanatory theories of Thibault and Kelly, and Richard Emerson.[1] The sense of rationality to be advanced, however, is a "procedural" one, and this is what makes that proposal important.

In the next two chapters I want to build two bridges between the detailed critiques of earlier sections and the outlining of a normative interpretation of exchange theory. What I hope to show is that there are more general sorts of

arguments motivating the pursuit of this alternative, and that for both normative and explanatory purposes an exchange-theoretic foundation has decided advantages over that provided by game theory.

The first of the two bridges is in the explanatory domain and concerns social power. What I hope to show in this chapter is that because Harsanyi's approach concentrates on the resolution of the situation the agents find themselves in, rather than on the agents themselves, his notion of social power is inherently static and restricted in scope. Harsanyi's conception does not treat how the interaction between two agents evolves, nor how the power balance between them changes as a result of the interaction, nor how an agent's strategic options include both the seeking out of specific coalitions external to the interaction at issue, and the altering of his values and beliefs. Even if we grant the game-theoretic foundation the ability to treat strategic power once we are able to assume what the situation is like, it is not enough, since in general one of the strategic options open to the agents is to simply change the situation.

A general explanatory theory incorporating the notion of social power must deal with such dynamic considerations as these. What seems to be required is a shift in focus away from the statically-defined "situation", and towards the agents and their ongoing relationship.[2] Richard Emerson's

theory of power-dependence relations takes this route.

In order to clarify the general arguments raised against game theory's conception of power I undertake a brief discussion of Emerson's theory. The major criticism I raise against that treatment is the present lack of a choice mechanism indicating in more detail how agents react to various power situations. The structure of the choice problem is all there, but Emerson backs away from giving specific details on how his theory sees an interaction's evolution taking place. This was no doubt intentional, since his concern was with the development of a set of sociological concepts. It is up to us to develop an economic choice rule on the same foundation.

The second bridge is in the normative domain and concerns more general objections against the theory of games. As the preceding discussions have illustrated, of particular concern is the adoption of the assumptions lying behind the normal form of representation as a characterization of the context of interaction. In Chapter VIII I raise general objections to the normal form, and show that the objections raised in the discussion of social power can be extended to argue for a change in our approach to the normative theory of rational interaction.

Of particular importance in this change in normative theory from a game-theoretic model to an exchange-theoretic model, is the shift to a procedural approach to rationality.

As has been indicated, this shift involves the recognition that human action and human interaction are rational, and that it is through understanding them as such that we learn both how to explain social behavior and about the normative issues raised by human interdependence.

Any normative interpretation of exchange theory, however, encounters two specific objections. First, the a priorist will object that the normative theory now has empirical content. Secondly, the individualist will object that Emerson's theory embodies a social-structural primitive. The first objection argues against any procedural notion of rationality; the second against such a notion coming from any but a psychological theory.

Although any full examination of these objections would take us far beyond the scope of this essay, in Chapter VIII I offer what I take to be plausibility arguments defending the exchange-theoretic approach against the assaults of both the reductionist and the a priorist.[3]

With this groundwork done in both the explanatory and the normative domains, Chapter IX presents an outline of the normative interpretation of the exchange approach to rational interaction. In that chapter I discuss in some detail the alternative focus for a theory of rational interaction, the "procedural" approach, and suggest further reasons why its adoption would lead to a fruitful research program on both the normative and the explanatory sides.

7.2. The Underlying Structure of Social Power Models

In Chapter V, I argued at some length that Goldman's theory of social power failed because he chose to focus initially on the model of a single agent acting in a "non-social situation" considered as a given. Whatever the merits of his basic analysis on other counts, as a starting point for a model of social power it proved to be his undoing.

Harsanyi, of course, did not encounter Goldman's problems. His theory of games provided him with the normal form of representation as a ready characterization of the context of interaction between agents. So long as the way the agents view their interdependence and the values they attach to the consequences of their actions can be modelled together by a game in normal form, Harsanyi has no trouble defining the context of interaction.

In the two-agent case, the as yet unexplored alternative to these is to focus on the agents themselves. Whereas in some cases the "situation" may be treated as an independent variable, as Goldman wants, and in some the normal form may be an adequate characterization of all the properties of the beliefs and utilities of the agents relevant to the interaction, as Harsanyi wants, in general any statically-conceived "situation" is merely a defined variable subject to change by choices made by the agents themselves. In general, that is, the theoretical framework suitable for an exploration of social power must go deeper than the superficial

characterization of static "situations" and must focus on the interdependence of rational agents and how this interdependence changes as the agents interact.

Now, there are two obvious criticisms of Harsanyi's model which he could easily overcome. Harsanyi knew that his approach did not model situations in which the agents happened not to satisfy the total knowledge assumption. However much we want to reject this assumption, it was acceptable to him since conceptual clarity could still be had even if explanatory scope suffered, and he was aiming for the former.

Harsanyi also knew there were some conventional exercises of power not convertible to the offering of conditional rewards and punishments, for example, the offering of unconditional bribes. But these, he contended, were strategically not as interesting, and could probably be treated as degenerate cases. Whether that is the case or not, this criticism would not do Harsanyi much damage.[4]

However, there is another sort of question which Harsanyi's approach is in principle unable to handle, and I want to argue that it is this inability that argues for the need to shift to the new focus for social power outlined above. This involves the raising of questions concerning the generation of the "situations" of concern to Harsanyi. That is, what are of concern in general are both the statically-conceived situations of the normal form sort, and

the generation of these as a result of strategic decisions by the agents. Now, it would seem clear that social power is involved in the generation of such situations no less than in their resolution. Such examples of interactional strategies as an agent's changing his utility function to avoid suffering unduly in a "forced" interaction, and his choosing to alter the set of strategy options when he feels the present set is too restrictive, are both evidence of this.

But if a general notion of social power can only be generated in a theoretical framework capable of treating the generation of situations as well as their static resolution, two crucial changes would seem to be required. First, we must admit into our analysis a "social-structural primitive", since the variables characterizing the "situation" and those characterizing the "agents" are now interdependent. The consideration of the dynamics of agent change seems to force this on us. How we are to understand such "social-structural primitives", however, is not yet clear.[5]

Secondly, in recognizing that the "situation" is merely a defined variable anyway, we might well be advised to shift our focus to the agents themselves, their interdependence, and the dynamics of the evolution of their relationship.

These two changes will now be examined and defended.

7.3. Social Structure and the Agents

I first want to dispose of a few of the usual psychological-reductionist arguments transformed to defend Harsanyi's treatment of social power. Although it is never clear exactly what the reductionist is arguing for, by considering a few illustrations of his arguments I hope to clarify the type of social-structure of concern to a theory of rational interaction.

The first argument runs roughly as follows. While it may be the case that the generation of the "situation" at issue is not explained by considering only the parameters of this situation, this does not mean that that generation could not be adequately explained by considering another situation in which the generation of the one at issue appears as one of the strategic options open to the agents. If that can be done, then there is no requirement for any "social-structural primitive" in explaining either what the agents do in the present situation or how that became the situation they found themselves in. Therefore, even if the generation of situations is part of what social power must explain, that can be handled without recourse to social-structural primitives. Everything is ultimately explainable in terms of individual actions by individual agents.

What this argument assumes, of course, is that the full explanation of the prior situation (and therefore the explanation of the generation of the present one) does not

itself require a social-structural primitive. Where I argue that it does, the reductionist will try to argue for its elimination on the grounds of a further retreat to a prior choice situation. But at whatever stage of the regress he chooses to stop, the reductionist will need to assume by fiat the explainability of that choice situation without a social structural primitive, and there my response will force the issue on.[6]

What this response does, of course, is simply to defuse the argument. It does not by itself show that social-structural primitives are necessary for an adequate treatment of social power.

Another argument the reductionist often mounts is simply to deny the existence of social structure. The charge becomes, "If you think there are ineliminable social variables, show me one and I'll show you how to eliminate it". I think this line of argument rests on a crucial confusion. It would seem that the reductionist arguing this way thinks that there is some theory-neutral way to define a "social phenomenon", and that if a psychological theory is available which adequately explains it, then by "something like" Occam's razor, no "social" theory can be justified.

Now, I think this argument can be disposed of without going into any detail on the issues of theory-neutral description or the use of Occam's razor in the explanatory domain, although both of these are pretty shaky grounds for

the reductionist's case. What I think is involved here is an implicit resort to the first argument discussed above. That is, the reductionist's "elimination" will amount to nothing more than an "explanation" which takes other interesting social-structure as background data, and we have already seen the folly of that line of argument.

Of course, there is another confusion lying behind this "elimination" argument. What the reductionist really wants to "eliminate" are supra-individual entities -- things like "the group", or "collective mind" or "the One" -- treated in the theory by primitive referring terms. But if this is his worry, he may relax his vigil. No such terms are needed in the alternative theory of social power to be presented here. The only primitive entities are individual agents, and they are treated as having something like utility and belief functions, and that is all. There are no ghostly presences to be exorcized.

But individual agents do interact; they are interdependent, and it is their interdependence that demands a social primitive. How that requirement shows itself is in the dynamics of interaction.

Consider, for example, an agent who changes his values because interaction with an agent he is dependent on for survival would be less costly to him if he did. To fully explain that phenomenon clearly requires a discussion of the agent's dependence on the other. But that dependence is a

Before proceeding to that discussion, however, I want to return to game theory as a normative theory, and present some more general arguments against that approach to rational interaction.

of two variables, although the exact connection between the three terms is an empirical matter. As Emerson put it:

> The dependence of actor A upon actor B (D_{AB}) is (1) directly proportional to A's motivational investment in goals mediated by B, and (2) inversely proportional to the availability of those goals to A outside of the A-B relation.[10]

In this definition "motivational investment" is a function of (1) the value of the rewards B mediates for A, and (2) the number of the different types of rewards B mediates for A. "Availability" is a function of (3) the number of relations which could be considered "alternatives" to the A-B relation as far as A is concerned, and (4) their respective value "comparison levels". Each of these four variables are fundamental properties of the primitive social-structural term, the "exchange relation".[11]

Each is examined in more depth in Chapter IX, where the connection between the "procedural" approach to rationality and these operant-theory-like notions is explored. For present purposes all we need consider are the two terms, "motivational investment" and "availability", and note that the latter is clearly a property of the social structure.

Emerson proceeds to define power in a manner reminiscent of Weber:[12]

> The power of actor A over actor B (P_{AB}) is the amount of resistance on the part of B which can be potentially overcome by A.[13]

In other words, P_{AB} is the level of potential cost A can induce for B, where the cost of a given transaction is

the value of reward foregone in making that transaction. Thus, power is seen to depend on the range of transactions undertaken as part of the exchange-relation, and on the availability of other sources of the "same" gratification.

For reciprocal power situations, A's power advantage may be defined as:

$$(7.1.)\quad \text{Power advantage} = P_{AB} - P_{BA}$$

since in such situations the above definition clearly attributes power to both agents.

Now, the connection between power and the fundamental property of dependence is stateable as the simple theorem:

$$(7.2.)\quad P_{AB} = D_{BA}$$

The proof of this theorem does involve one fairly weak empirical proposition, however. Emerson labels this Proposition 1. As he shows,[14] in addition to the definitions we need to assume that the probabilities of an agent's performing alternative acts to achieve some given goal vary directly with the probabilities that those acts provide him with that goal.[15] Of course, as the agent learns which act is most likely to provide him with the desired goal, he may choose to perform that act exclusively, but this claim involves a larger empirical assertion, and we can do without it.

Now it is argued in more detail in Chapter IX that this

"empirical assumption" embodies a sense of "minimal rationality". Emerson does not seem to recognize this (or else he thinks it unimportant), and never explores the possibility that a rational choice rule could operate within his theory. It is that extension I want to explore as providing a procedural approach to the normative question. For present purposes it is sufficient to note that in generating his theory of power, Emerson has (perhaps unconsciously) built in a normative model of the agent. What is left to be done is to extend this model so as to answer a number of more specific sorts of questions Emerson's theory left unanswered, on the grounds that the "exact relationship" was an empirical matter.[16]

In general social situations, of course, it is reciprocal power relations that are important. When both agents are dependent on one another, the state of "balance" or "imbalance" of their relation can be most revealing. Following the definitions just given, a balanced relation is clearly one where:

$$(7.3.)\quad \begin{array}{ccc} P_{AB} & = & D_{BA} \\ \| & & \| \\ P_{BA} & = & D_{AB} \end{array}$$

while an unbalanced relation is one where:

$$(7.4.)\quad P_{AB} = D_{BA}$$
$$\vee \qquad\quad \vee$$
$$P_{BA} = D_{AB}$$

Now, if a power imbalance occurs, the stronger member can overcome some degree of resistance and attain goals mediated by the weaker which he could not attain if power was balanced. That is, he can force him to accept some "cost" as defined above. For example, if a lonely but puritanical girl is asked on a date, she may assent to moderate sexual advances owing to her dependence on her date's attention in overcoming her loneliness. But such an exchange is not without cost to the weaker member, since she may suffer some degree of guilt owing to her attitude towards sexual activity. In such a situation she might be able to reduce those costs by, for example, discarding her puritan values. In general, such cost reduction processes can occur in any relation where costs are "anchored in modifiable values and attitudes".[17]

But these sorts of "cost reduction processes" do not necessarily alter the state of balance of the relation, even if they do tend to strengthen and stabilize it through decreasing the number of equally acceptable alternatives. Other operations lead to a change in balance.

Given the conditions of equation 7.4, there are four variables affecting the state of balance. Each of these provides a generic type of balancing operation:[18]

1) A decrease in the value of what A provides for B.

2) The cultivation of alternative sources by B for the values mediated by A.

3) An increase in the motivational investment of A in the goals mediated by B.

4) A reduction in the alternative sources available to A for the goals mediated by B.

The first operation we may label "motivational withdrawal"; the second leads to the extension of the "power network"; the third is exemplified by the emergence of "social status"; the fourth involves "coalition formation" and the development of group norms.

To explore these four balancing operations in any depth would take us too far afield, but two things stand out as worthy of note.[19] First, the combined notion of power-dependence is able to provide a foundation for the development of the usual range of social-structural terms like group, norm and status. In fact, the only social structural primitive needed for the whole theoretical framework is the notion of an "exchange-relation", and its key property is dependence. All other theoretical concepts are introducible on this foundation.

Secondly, it should be noted that what each of these balancing operations involves is a change in "situation" -- just the dimension lacking in Harsanyi's theory. Of course it has never been denied and it may well be the case that Harsanyi's game-theoretic approach could reconstruct many of the specific transactions between agents. But what the

present model gives us over and above that is a classification of four generic types of operations whereby situations evolve. Since balancing operations are clearly dynamic strategies, they enrich the problem of strategic choice by opening up the connections between choice and the full range of social-structural development.[20]

But what Emerson has not done is show us which cost reduction or balancing strategy will be chosen in a given situation. He has given us a rich theoretical structure in which the options and the relevant considerations are laid out, but he has not given us a general choice mechanism. He has, of course, made a few suggestions and it is these I develop in Chapter IX.

However, there is one important mechanism he does explore, and since it pertains to the use of power I will examine it here. This mechanism is summed up by the aphorism:

> To have a power advantage is to use it; and to useit is to lose it.[21]

Consider two interdependent agents, A and B, each having a behavioral repetoire containing items which either could be, or already are, found gratifying by the other. Now, some degree of power is used in all social exchange since both agents employ items in their behavioral repetoire (their power base) in order to gain access to items in the other's behavioral repetoire. But an important question concerns when an agent's use of power may be said to be increasing over a series of transactions, since this will

give some insight into the altering of the state of balance of the exchange relation.

Emerson suggests that:

> [Agent] A's use of power is said to be increasing across a series of transactions if (a) B's costs increase through additional ...[acts] by B, or (b) if A's costs decrease through decreased [use of resources] ...without decreasing rewards to A.[22]

Now Emerson suggests that owing to the key empirical proposition he labels Proposition 1 and which I have suggested embodies a normative model of the agent:

> ...if A's power is greater than B's current costs and if B has additional resources, A's use of power will increase, cutting further into B's resources, until its use if offset by incurred or anticipated costs to A.[23]

If we restrict the discussion to costs anticipated or incurred within the A;B exchange relation, the limit on A's use of power is his own dependence in the relation. Thus, in a balanced relation, "increase" in the use of power is unlikely, while in an unbalanced relation, A's use of power will increase as a function of power advantage.

That is, by the key proposition in Emerson's theory, the one I argue embodies a minimal normative model of behavior, to have a power advantage, is to use it. Exactly when that advantage is used, however, is another matter.

But one of the effects of an agent's use of a power advantage is either to increase the rewards he obtains from the relation or to decrease his costs. In either case, his dependence on the relation increases since fewer equivalently

rewarding alternatives will be available which provide the same range of transactions. Moreover, since the other agent may be suffering greater costs, his range of "equivalent" alternatives will (in general) increase, and so his dependence on the relation decreases. In general, then, the tendency is for the dependence of the agents on the exchange relation to become equal. In short, to use a power advantage, is to lose it.

Now these very basic considerations of the use of power need to be made much more elaborate before Emerson's approach to power can challenge Harsanyi's as far as the relative strengths of their strategic theories are concerned. Emerson has not offered a choice mechanism capable of treating any specific decision on the part of an agent party to a power-dependence relation, and so has not even broached the prescriptive problem.

However, he has placed interdependent decision-making in a compelling explanatory framework, and has left the development of a more sophisticated normative theory within that framework an open problem. It is that problem I turn to in Chapter IX.

7.5. Conclusion

This chapter concludes the discussion of social power. What I have tried to show is that a theory of power needs to be built on a normative foundation since the exercise of power and the response to power-based overtures depend on strategic choices by the agents. Goldman's theory did not have such a normative model, and floundered on many key issues as a result.

On the other hand, Harsanyi's game-theoretic approach was too restrictive. Most importantly his theory could not handle important dynamic considerations. The inadequacy of his normative model as a normative model, of course, was a further problem.

Emerson's explanatory theory of social exchange, reviewed only briefly in this chapter, overcame the problems forced on Harsanyi by the combination of game theory's static approach to interaction and his own dual-interpretation approach to rationality. But Emerson's treatment of power was seen to have at base only a minimal rational model. This made his handling of social power too general for specific explanatory work, and totally inadequate as a general normative model.

The next important problem in exploring the potential of the exchange-theoretic approach to explanatory relations like social power, then, is to extend its normative foundation. This is the topic of Chapter IX.

Before proceeding to that discussion, however, I want to return to game theory as a normative theory, and present some more general arguments against that approach to rational interaction.

FOOTNOTES

1. See especially Thibault and Kelly (1959), Emerson (1962), (1972a) and (1972b).

2. As will become clear, the "relation" between two agents will be treated as a social-structural primitive. Properties of the "relation" are not reducible to properties of just one agent or the other, but are attributable to the interdependence of such individual properties. Power, for example, is not solely a function of one agent's ability to manipulate the flow of goods to another, but is also a function of the other agent's dependence on those goods and on this particular source of them.

3. As to psychological reductionism in sociology, Richard Emerson has defended exchange theory at much greater length than I can here. (See especially his (1969).) His main argument, however, is similar to the tack I adopt in that he contends that what psychology, in particular operant principles, cannot tell us is how any particular stimulus environment (to which individual organisms react) came to be organized the way it was. Social-structural concepts like coalition, status, norm and so on, are needed to do that. See, particularly, the discussion in (1969), p.403, where he describes his strategy of "construction" on a psychological foundation as contrasted with "reduction" to it.

4. Both of these arguments are acceptable given the task that Harsanyi set himself. Neither restriction arises in the approach to be explored here.

5. The connection between, on the one hand, exchange theory and classical political-economics, and on the other, game theory and neo-classical microeconomics, is obvious. In gross terms, both of the latter treat the beliefs and values of the agent as givens. In neo-classical theory, the model becomes stylized in the atomic form of "rational economic man". In his (1972) Vickers summarized the concept very neatly:

> He was supposed to know what he wanted and to know what disposal of his resources would best satsify his wants. He had no economic wants which

could not be satisfied through the market. All he needed was access to a market of goods and services which, being "free", would be responsive to his wants and to a labour market which, being also "free", would be responsive to what he had to offer. ... I stress that the economic man was an "atomic" concept, a separate entity, related to his fellows only through the working of the market. His wants were his own affair; they could and should be taken for granted.

For a more detailed discussion of the classical versus the neo-classical dispute, Hollis and Nell's Rational Economic Man (1975) and Adolph Lowe's, On Economic Knowledge (1965).

6. This response is particularly telling when we restrict ourselves to the domain of human intentional action and its social environment. I am not concerned here with the "natural" environment and whether or not explanations in that sphere are adequate without such collective entities as "fields" and "mass distributions".

7. In Ryan, The Philosophy of Social Explanation (1973).

8. See Emerson (1962), (1972a) and (1972b).

9. Emerson (1962), p. 32.

10. Ibid., p. 32.

11. Each of these terms is defined in great detail in Emerson (1972a). (See especially p. 57.) In fact, the objective of that paper was really to build these concepts on an operant-theory foundation. As indicated they converge on the notion of dependence, which is the key property of exchange relations for Emerson's social exchange theory.

12. Weber's definition has guided most of the work on social power in the last thirty years. "Power is the probability that one actor within a social relationship will be in a position to carry out his own will despite resistance" (Weber, 1947, p. 152).

13. Emerson (1962), p. 32.

14. Emerson, (1972b), p. 64.

15. This is a paraphrase of Emerson's Proposition 1, from his (1972a), p. 46.

16. This interpretation was first suggested to me by Jim Leach. See his (1975) for a discussion of some of the issues raised by this dissertation.

17. The example is close to one given by Emerson (1962), p. 34. The generalization is also his (p. 35).

18. Ibid., p. 35 and (1972b), p. 67-68.

19. Large sections of Emerson's (1962) and (1972b) are devoted to exploring the consequences of each of these balancing operations.

20. In his (1972b), p. 60, Emerson points out that the concept of the exchange relation between actors A and B (symbolized (Ax_i; By_k) with x_i and y_k variable behaviors) focuses attention on fixed A and B with variable transactions across the lifetime of the relation, while market economics assumes a fixed transaction (x,y) with variable actors. How this linkage might be developed into a unified theory is a very interesting problem for future research. (See, for example, Leach (1975).

21. Emerson (1972b), p. 67. Note that Emerson is here concerned with exchange relations in which transactions continue to occur. Should no further interaction take place, of course, it is clear that power could not be employed.

22. Ibid., p. 65.

23. Ibid., p. 65.

CHAPTER VIII

EXTENSION OF THE CRITICISM OF GAME THEORY

8.1. Introduction

The criticisms raised against Game Theory thus far in the dissertation have been aimed directly at Harsanyi's tracing procedure and his bargaining model. It has been argued that for a number of reasons both proposals fail two of three criteria of adequacy Harsanyi is committed to satisfying.

Throughout the discussion, however, there has been an underlying suggestion that the major fault lies not so much in the details of Harsanyi's models as in the assumptions lying behind the traditional approach to game theory. I now want to extend some of the criticisms levelled at Harsanyi's programs to show more clearly that the normal form of representation ought to be rejected as the foundation for the theory of rational interaction, and that any alternative leads in the direction of a procedural notion of rationality.

In Chapter II it was shown that there are only two ways in which the interdependent choice problem in normal form can be "solved":

1) Show that some sort of rationally defensible iterative reasoning regress converges onto an equilibrium point, or

2) Show that some special combination of strategies can be taken as the solution on the grounds that they satisfy intuitively acceptable axioms which "short circuit" the regress.[1]

In general, game theorists have been preoccupied with the latter approach. Because it has led to distinct treatments of different types of games I call it the "patchwork" approach.[2]

In this chapter I outline a number of more general arguments against the normative use of game theory. First, I suggest that its apparent "successes" notwithstanding, the patchwork, axiomatic approach to generating solution concepts for game theory is unsatisfactory. It is not so much that formal axiomatics is wrongheaded, as it is that what we are trying to axiomatize has not really been thought through. Of course, what I want to suggest is that we should first generate an interesting procedural notion of rationality, and then see what formalization can tell us.[3]

Secondly, I argue that as an alternative a model based on iterative reasoning is irredeemably constrained by the normal form. There is not enough leeway in the way it models the expectations of the reasoner to let the iteration get off the ground.

Finally, I outline overriding philosophical reasons why the normal form might be abandoned as the foundation for a theory of rational interaction, and show how a

procedural, exchange-theoretic approach to normative theory relates to these objections.

8.2. The Patchwork Approach

There are any number of philosophical reasons why an all-encompassing general theory of any sort is preferable to a theory consisting of a set of special axiomizations.[4] As it now stands, traditional game theory has only given us a few of the latter sorts of sub-theories. The most notable among these have been von-Neumann-Morgenstern Zero-Sum theory, the trivial application of the "Sure-thing Principle" to some non-cooperative games, and various Nash-like theories for cooperative games.[5]

However, there are three reasons why even these paradigm "successes" are open to question as bulwarks of the patchwork approach. The first of these is the argument that all they really amount to is axiomatic reductions of special cases of a more general iterative reasoning model, and that our attention ought to be focused on the development of that general model. Since an iterative reasoning approach would be more general, it is preferable as the foundation for a theory of rational interaction. An axiomatization of the general model, of course, might still prove to be very revealing.

We should also note that if psychological models of practical reasoning are relevant to the sorts of normative

issues raised by game theory, then some sort of iterative reasoning approach is prima facie more plausible anyway, since we do seem to reason sequentially. In particular, if we relax the perfect information assumption, it would seem clear that in non-cooperative games at least, we do try to "out-think" our opponents, and iterative reasoning is ideally suited to capture that phenomenon.

Secondly, game theory can be criticized for having been overly committed to an extreme sort of reductionism. Since zero-sum games are the most amenable to mathematical treatment the objective has often been to reduce other sorts of interaction to that ideal format, whatever the obvious dissimilarities.[6] No one really disagrees that one of the results of this reductionism has been that the explanatory value of mixed-motive theory has suffered, but I also want to suggest that this failure is prima facie evidence that its normative worth has also suffered. The general issue is too involved to be pursued here.[7]

This criticism of the reductionism inherent in traditional game theory has been voiced by Thomas Schelling, who has long argued for a theory of strategy having the general mixed-motive game as its primary focus.[8] Schelling's examples of pure coordination games illustrate what he feels this bias has brought game theory to neglect. There are strategic elements, he argues, in the way the players develop a "shared-perception" of the situation, and if we explore

these, it can lead to an understanding of how they go about solving their interactional problems.

One of the reasons why traditional game theory ignores this sort of problem, of course, is because it concerns the dynamics of situation generation while the normal form deals with situation resolution. Another is because the assumption has always been made that it involves no interesting strategic elements. Since the first of these defences has already been discussed, the second will be treated here.

What the perceptive game theorist would argue, of course, is not that there are no strategic considerations in the development of shared perceptions of some situation, but that if there are, then by definition of the normal form they must either already be incorporated in the game-model as it is presented, or we are simply disagreeing over what the situation is that that normal representation is supposed to be modelling. Since the whole function of the normal form is to represent strategic options, there can be no other alternative.

What this reply misses is that the issue raised by Schelling concerns how the agents are to be modelled and not just what the choices they confront happen to be. The strategic matters of concern to Schelling are really epistemological strategies employed in an interactional context. Part of an agent's perception of the situation is what he "should believe" about his opponent. But that is a

function of what the other "should believe" about him. (This looks very much like a reciprocal reasoning regress at the level of situation perception.) However, the starting point for each agent's reasonings on this matter is not restricted to a normal form, and can be anything he knows about his opposite number. (Of course, the focus they are looking for can concern any belief of either agent; it need not be simply what one of the agent's strategic choice will be.)[9]

Schelling's theory fits rather neatly into the procedural model of rationality discussed in Chapter IX, and so I will not pursue it further here. The short reply to the game theorist is simply that the development of "mutual perceptions" concerning a given mixed-motive game involves another sort of interdependence, and this can be approached iteratively. Where the model differs from game theory is in its denial of the assumptions defining what the rational expectations of rational players are regarding their opponents <u>given</u> a context of interaction. These expectations are worked out by rational players during the interaction itself, and the focus for the solution of the coordination problem can be anything in the context of the interaction or in the psychologies of the agents.

The third general sort of criticism of the traditional patchwork approach I want to consider is that its paradigmatic "successes" might not really be all that successful after all. Of course, there can be no serious doubting of the

formal properties of the "solutions" it has proposed, but whether or not they satisfy normative desiderata is still very much an issue.

The first such paradigm is von-Neumann and Morgenstern's Minimax Theory and its definition of a value for all two-person zero-sum games. The intuitive appeal of their model is still very great, of course, but it has recently come under criticism on a new front. The details of the arguments raised against it are far beyond the scope of this essay, and I can only refer the reader to McClennen's excellent discussions for a powerful presentation of the case against the Minimax Theory.[10] His fundamental argument is similar to the one I have employed against both Harsanyi's use of an "estimator" giving other than best-reply estimates, and against the assumption that threats are credible to rational players. Quite simply, maximizers have no reason to play according to the proferred strategy. As it happens, McClennen feels that the maximization of expected utility choice rule is what must give way in this confrontation. His main argument, though, supports the resolution I have alluded to here.

A second paradigm offered as proof that the patchwork approach to normal games is a viable program for generating normative theory is the so-called "sure-thing principle". Surely the best advice in non-cooperative games is to choose that strategy, if one is available, giving at least as high

a payoff as any alternative for each choice open to the opponent. So argues the game theorist, and the applicability of the sure-thing principle to normal games seems to vindicate the latter as a viable model.

Once more, however, the weight of the intuitive evidence is not all on one side, and if the verdict is reversed, either the normal form is missing something, or the sure-thing principle is less "sure" than it appears.

Three issues stand out as giving most pause to sure-thing theorists:

1) Nigel Howard's proof that in certain mutually-ordinal games the sure-thing user is doing himself a disservice.[11]

2) The long-standing problem of the rationality of the "double-cross" resolution of the prisoner's dilemma game.[12]

3) The very intriguing issues raised by Newcomb's paradox.[13]

Again, each of these issues would take us too far afield, and I can merely register the observation that applied to games in normal form, the sure-thing principle does not lead to uniformly intuitive results.

The third paradigm success of the patchwork approach to games in normal form is the Nash proof of the existence and uniqueness of the solution to two-person bargaining games with given disagreement payoffs. There is no doubt that this is an interesting result, but whether it vindicates the patchwork approach and the normal form is another matter. Nevertheless, it is worth noting that Harsanyi's demonstration

of the equivalence of Nash's solution and Zeuthan's iterative approach, at least leaves it open whether game theory has given us any new insight. If the results of the two approaches had been other than identical, one wonders which Harsanyi would have chosen.[14]

What I hope these brief remarks indicate is that the concerted attempt of the past thirty years in game theory to prove normatively interesting theorems about the normal form has been less than spectacularly successful, and that a return to an iterative reasoning model is not too unreasonable. Of course, it is against this background that Harsanyi's return to the naive reciprocal-reasoning regress in his tracing procedure stands out in greatest relief.

8.3. The Iterative Approach to Normal Form Games

Harsanyi's concern with providing psychological interpretations for his models seems to have driven him to avoid the usual axiomatic approach to game theory.[15] In turning to the iterative reasoning regress with his tracing procedure, he focused attention directly on the three key elements of any model of practical reasoning:

1) What data are assumed to be available?

2) What sort of algorithm is turned loose on this data?

3) When does the rational player stop computing and start acting?

I want to argue that the way the normal form is defined, it

is impossible to answer these three fundamental questions in any happy way.

First, the normal form requires that the only relevant data available to the players at the "start" of the reasoning regress comprise the game vector itself. The players and the context of their interaction are reduced to a set of fixed strategy options and fixed utility assignments for each matching of strategy choices. Of course, as a reconstruction of what the strategic options are at any instant of choice, there is little to quarrel with in this model. But the use of a static reconstruction as opposed to a dynamic model is only justified if it leads to normative insight. If the really interesting normative issues concern how rational players arrive at such an end point (and it may be that it is rational to arrive only at certain sorts of such "normal" games) then static reconstruction would seem to be insufficient.

The iterative model provides the escape from the problems raised by static reconstruction but only if it rejects the assumption that the final normal form must represent the starting point for the regress. The only thing barring this rejection is the power of the normal form. But the power of the normal form as a reconstructive device does not entail that it must also be the starting point for the iterative algorithm. The function of the iterative algorithm might be not to "solve" each given game, but rather

to arrive at ones that are solvable by other means.

In short, it is only if we assume that nothing about the agent's perception of their interaction changes during their iterative reasoning that the reconstructive power of the normal form means that any iterative reasoning algorithm need adopt it as a definition of its starting point. If we deny the antecedent, and I want to do that, then the way is clear for a more general sort of iterative model of practical reasoning in interdependent contexts with other sorts of objectives than simply "solve this game".

The normal form also requires that any "stopping" point for an iterative procedure be an "equilibrium point" in some sense. In the case of non-cooperative games it was an equilibrium point of the game. In the case of cooperative games (if an iterative model were developed to duplicate Harsanyi's approach) the end point would have to represent an equilibrium among the set of all coalition agreements. In both cases it was pointed out that equilibrium itself has no special appeal to any rational reasoner. Now, if the arguments raised in previous chapters are valid, then the generation of equilibrium-ending algorithms has received a serious setback. But independent of this claim is the objection that the normal form's requirement that any acceptable iterative algorithm end with an equilibrium condition begs the important empirical question as to what sort of iterative algorithm does in fact model psychological

processes. Equilibrium may be relevant, but only empirical theory can inform us of that.

This objection strikes at the heart of Harsanyi's dual-interpretation approach. Whenever interesting empirical questions are raised Harsanyi opts for an a priori resolution on "normative" grounds. However, he still wants the resulting theory to have "positive" application, and so he has to fudge the result. One way he does this is to suggest that: "Of course, for empirical applications 'some' of these assumptions may have to be relaxed". Other economists suggest that although no individual empirical situation can be modelled on such a strictly a priori basis, deviations from such a model are statistically random, and so the model still serves positive purposes.

But whatever moves are made at this point, the central conflict remains. The dual-interpretation approach assumes that people act rationally, and so on the one hand plans to use a normative model for positive purposes, but on the other hand refuses to study how people do act in order to generate the normative model in the first place. A "procedural" approach, of course, would base the normative model on such studies, and therein lies the difference between the two.

The third and perhaps most important question is whether any iterative reasoning algorithm can be applied to games in normal form even given a definition of the starting

point and the requirement of convergence to an equilibrium point. This issue can be most clearly faced in non-cooperative games, where the case against any form of iterative reasoning rests on the tension between the two dimensions of rationality implicit in the notion of a "rational estimate of a rational player's likely strategy".

The naive iterative Bayesian approach can be considered as an extreme attempt to resolve this tension. In that approach the nature of the iteration was such that at any point in the regress the estimate provided is "rational" just because it is a true reflection of the opponent's dispositions at the previous iteration. Each estimate simply is the best-reply selection of the rational opponent at the previous iteration, and so is clearly a "rational" estimate so long as no further computation is allowed.

The tracing procedure moved away from this justification of the estimate on the basis of its being certain at that iteration, and towards the other extreme of total uncertainty.

The estimate the tracing procedure provides is "rational" in the sense that it alters an extremal state of ignorance in some "optimal" fashion. That is, even if granted the rationality of the equipartition assumption across each opponent's "assessments", Harsanyi must still show that the way this initial ignorance is removed as the iteration proceeds is a "rational" adaptation to an improved state of knowledge.

Now, after the first iteration of the tracing procedure each player's newly acquired knowledge consists of the following rational beliefs:

1) that the "prior" rationally reflects the dispositions of the rational opponents;

2) that "my own" best reply is as computed; and

3) that the opponents have followed exactly the same format in their reasoning, i.e. they have assumed equipartition across my "assessments", have computed the prior, and have computed their own best-reply strategies.

The disposition of each player after this initial iteration is to play a best reply, and all players can compute the one each opponent is disposed to play. Now, as I argued at length in Chapter IV there is no apparent reason why rational players with this knowledge should be considered "ignorant" about each opponent's likely behavior, whether further iterations are allowed, or whether all players must choose a strategy without any more computation. Whatever ignorance they started with was removed by the universal equipartition assumption (assumed valid for the purposes of discussion).

The conclusion we must reach is that the tracing procedure estimator cannot be justified as providing a "rational" removal of ignorance. Therefore, as an alternative to the naive regress, it is unacceptable.

Now, for these same conclusions to apply to <u>any</u> iterative algorithm applicable to non-cooperative "games", it is sufficient to show that given formal encoding of total or partial ignorance, the transparency of reasoning

assumption implicit in the normal form removes in one step of the iteration whatever that initial state of ignorance is assumed to be. Once that ignorance is removed, of course, nothing stops the rational reasoners from entering upon the naive regress with all its problems except the ad hoc stipulation that that would not be "rational". Given the analysis of the previous paragraph applied to total ignorance, I think the conclusion is obvious.

The only alternative for normal-form games, then, would appear to be to spurn the equipartition assumption and all similar alternatives for handling ignorance. But as I have argued, this results in either an infinite backwards regress of the naive sort, or in a denial of mutual rationality. Neither is an acceptable resolution of the game in normal form, and so no iterative algorithm can work when applied to such games.

What I have tried to show in this section is that the attempt to develop an iterative reasoning model to be applied to games in normal form is constantly stymied by the assumptions built into that representation. Therefore, if an iterative model is to be preferred to the axiomatic approach, that model must be developed in conjunction with a weaking of some of the assumptions lying behind the normal form. How this might fruitfully be done will be the subject of concern for the remainder of this dissertation. That such a weakening is tantamount to the generation of a

procedural notion of rationality will, I think, become clear. The only serious question that remains is whether philosophers will recognize this and turn directly to appropriate explanatory theories for their insights or continue under the illusion that a priori argument is all that matters.

8.4. Some Philosophical Prejudices Countered

Many of the arguments of earlier sections have amounted to criticisms of the normal form on the grounds that there are things we should be able to do with it that we cannot. It is eminently suitable for reconstruction of choice situations, but there appears to be no way to have it represent the dynamics of situation development. We might want a normative theory to do that. It assumes that the players of the game have fixed expectations about the dispositions and beliefs of their opponents. It might turn out that both the expectations and the dispositions change during rational interaction. It is formulated so as to facilitate mathematical treatment. The concern of normative theory is also conceptual and theoretical development. Finally, and more generally, it is unable to treat any properties of the agents' utility and belief functions other than those that can be cashed out in terms of strategy options and utility assignments to expected consequences. A normative theory of choice might have to be concerned with a model of the agent having greater structure.

Of course, any defender of the use of game theory as a normative tool has a ready response to this form of criticism. What justification, he asks, is there for loading down the normative theory with all the structure needed to handle these sorts of questions? To be sure, these issues and many more must be faced if game theory is ever to have direct explanatory application in the social sciences, but the objective in this branch of normative theorizing is to clarify the nature of rational choice in interdependent contexts, not to explain how people behave. To do the former we need make only those assumptions about the nature of choice and the nature of contexts of interaction that are necessary to lead to a general statement of the question "What is rational?" In its generality the normal form gives us such a model. The set of questions it leads to is not exhaustive, to be sure, nor are they easily answered. But when answers are arrived at they are normatively informative and we can ask for no more than that.

Such a reasonable-sounding reply unfortunately misses the main thrust of the criticism. The question is really whether the assumptions lying behind the normal form are the best way to achieve normative insight in these matters. The above response suggests only that because it is so general in its ability to reconstruct certain forms of choice situations, the normal form must be the best. That is, any other model will only be a special case. But in arguing

this way the response misses the obvious reply that what is really needed is a model which frames the same issues that the normal form does, but does more as well. In that way the normal form becomes a (perhaps not very interesting) special case.

Now, there are any number of ways such a more general framework could be developed. Non-Archimedian representations of preference might lead to interesting changes in the way we represent interdependent contexts.[16] So might a theory unifying epistemic logic and strategic reasoning.[17] I have no doubt that a theory of rational value-revision would also add much to a theory of rational interaction.[18] But the exchange-theoretic model also strikes me as being a fruitful line to explore, even if it is not at the present time as formalized as these other options.[19] What is needed to make this a viable option is an argument to show that there is a normative foundation to exchange theory. I attempt to do this in Chapter IX.

But there is one other objection that a game theorist like Harsanyi might raise even if that foundation could be generated for exchange theory. This objection is the last gasp of psychological reductionism.

The argument runs as follows. Even if it is accepted that social-structural primitives are needed in explanatory theories, and that the "procedural" approach which bases normative theorizing about rational action on appropriate

positive theory has validity, it is still counter intuitive to have a normative theory about human action depend on ineliminable social structure. It is the agent which is rational, and so surely the theory explicating that concept need make no essential reference to social structure.

There are two lines of response to this question. The first simply points out that once it is accepted that the best appropriate positive theory is to be used to generate a procedural notion of rationality, then the defeat of psychological reductionist arguments on the explanatory side ends the matter. But since this response may be felt to beg the meaning of "best appropriate", a second more detailed answer is called for.

What is really at issue here is not whether rationality is a property of the agent, but whether the other properties by virtue of which the agent is rational can be explicated without reference to social structure. That is, are an agent's beliefs, values and dispostions primitive individualistic properties, or does a full explication of them and their properties require social reference? If so, then not only is the concept of "an agent" infused with social-structural content, but then, so too, is the concept of rationality. In short, both who we are and our rationality are conceptually linked to our social environment.

Of course, the key to this more elaborate response is the move to make beliefs, values and some dispositions not

only causal but conceptual functions of the social environment. But this linkage is already established in the positive theory of social action (exchange theory) adopted as a premise in this discussion. Therefore the argument shifts back into the explanatory domain where by assumption reductionism has been effectively countered.

8.5. Conclusion

In this chapter I have offered some general arguments against game theory, the normal form, and the dual-interpretation approach to rational theory. The arguments against the patchwork axiomatic approach in game theory point to the need to develop an iterative model. The arguments against such an iterative model applied to games in normal form point to the need to develop an appropriate weakening of the a priori (reconstructionist) structure of the normal form. But this in turn suggests that we look to positive theory for a sense of direction, while if we do adopt that strategy, we find that the dual-interpretation approach has lost much of its force. The next step is to show how an appropriate positive theory can be adapted to serve normative ends without falling prey to the easy criticism, "But that 'adaptation' just applies a priori normative principles of optimality within the framework of a particular positive theory of human action". The key to avoiding this criticism, of course, is that the sense of rationality to be

developed, must be a procedural one. What must be shown is that human action can be conceived of as rational from the outset. With this positive program underway, we can then get on to the development of a normative theory which explicates why the theory's conception of action is a "rational" conception.

That we do wind up applying prior normative insight in this way does not detract from the major thrust of the procedural approach, however. Some ideal of what normative concepts might be like is clearly going to be required if any distinction between normative theory and positive theory is to be maintained, and it has never been denied that that distinction is important. But what the particular normative theory is like will still be a function of the positive theory on which it is (partially) based, and that is where the difference lies between the procedural and the dual-interpretation approaches.

This treatment of the connection between normative and positive theories allows for influence to be felt both ways, with gradual changes both in our explanatory understanding of action and in our normative understanding of rationality. But it does this without denying the distinction between the two.

The next chapter explores how this feedback works in practice.

FOOTNOTES

1. See Chapter II, section 2.1.

2. Part of the "patchiness" of game theory derives from the usual typology of games. The distinction between zero-sum (or constant-sum) and non-zero-sum (or mixed-motive) games, and cooperative and non-cooperative games reflect the tendency of game theorists to deal with extremal constraints in their modelling of situations of interaction: communication is either impossible or it is assumed to be universal and error free; collusion is either disallowed or it is backed up by enforced agreements; side payments are either disallowed, or we are to assume transferable utility, etc. Why the normative questions of rational interaction were first raised in such a piecemeal fashion is an interesting historical question I cannot explore here.

3. I find this a general failing of the formalist approach. It is almost invariably applied to an issue too early, and tends to obscure important conceptual matters.

4. A few of these are: conceptual connections are more highly developed; contact with potential falsifiers is maximized; and problematic interdependencies have not been ignored.

5. For details on these see Luce and Raiffa (1957). In looking at these three cases I do not underestimate the work done to date in game theory. However, I do think these stand out as the greatest achievements of game theory, and if they prove to be less than exemplary for normative purposes, game theory has lost its proudest successes.

6. John Nash was noted for this. See, for example, Luce and Raiffa (1957). Of course, the view of game theory as a theory of conflict resolution is partly to blame here. The shift to calling the objective a theory of "rational interaction" is intended to underscore the neglect paid by game theory to other aspects of interdependence.

9.3. Social Psychology and Exchange Theory

The theoretical framework of social exchange was developed during the past few decades to provide a conceptual structure capable of dealing with persistent problems in social psychology. The major contributors to this development have been George Homans,[4] Peter Blau,[5] John Thibault, Harold Kelly,[6] and Richard Emerson.[7]

The task these "theorists" confronted was the development of an "approach", "viewpoint", or "conceptual structure",[8] which could bridge the gap between psychological theories of behavior and small group theory, where the latter was viewed as the foundation for the theory of social structure. (Most saw this "bridge" resting on both sides of the gap, and so exchange theory is not essentially "reductionist" in spirit. George Homans, of course, represents the major exception to this rule.[9])

This chapter explores one relatively small, but important, contribution to this ongoing development -- the "structure" imposed on the exchange-theoretic framework by Richard Emerson. The major reason why Emerson's theory is chosen is that in two brief papers he has laid out the whole exchange-theoretic framework, from its operant-theoretic foundation, to the start of a theory of exchange networks, including status emergence, norm development, group formation, and so on.

In choosing Emerson, however, I do not want to give the

16. An Archimedean condition on a measure function is a second order axiom to the effect that standard sequences bounded from above and below are finite. Its basic function is to allow quantities to be compared. If it is dispensed with, certain sorts of preference non-comparability would be allowed. See, for example, Louis Narens, "Measurement with Archimedean Axioms" (1974) and R. Duncan Luce, "Conjoint Measurement" (forthcoming).

17. What I am thinking of here is a general theory of rationality treating choice and belief in the same framework. The Bayesians have tried to do this, of course, but have not had any great success in interactive contexts.

18. I have no knowledge of anybody working on this problem except learning theorists. See, for example, Emerson (1972a).

19. More importantly, the other revisions are only formal manipulations unless informed by something like exchange theory. In that way, an exchange-theoretic-model foundation would provide a rationale for the pursuit of various formal puzzles.

CHAPTER IX

A PROCEDURAL APPROACH BASED ON EXCHANGE THEORY

9.1. Introduction

This chapter ties together a number of threads left dangling in earlier sections of the dissertation. The central theme is the development of a normative interpretation, along "procedural" lines, of Emerson's version of exchange theory.[1]

The objective, though, cannot be simply to show how a cognitive model can be extracted from his framework, since, given the incomplete nature of most of his empirical propositions, that would be premature, if not impossible.

As a result, the conclusions drawn regarding the use of Emerson's framework for normative purposes are tentative and just as incomplete as Emerson's own conclusions.

In working towards this end, I take the following steps. First, the notion of "procedural" rationality is clarified. Then I show how Emerson's operant-theory-based approach to exchange theory provides a plausible framework for the development of such a model. As in the discussion of his theory of social power, the conclusion of this analysis is that the structure for the normative theory is present, but more empirical detail is needed in his theory before an adequate "choice rule" can be defined in cognitive terms.

Finally, some of the problems raised by this approach to a theory of rational interaction are discussed, and I indicate the work that lies ahead.

The conclusion, then, is that while a "procedural" notion of rationality underlies Emerson's theory of social exchange, further insight into the normative problem requires an elaboration of that explanatory framework.

9.2. The Procedural Approach

This section explores in more detail what has been suggested is a promising but underexplored approach to normative theorizing. I labelled this the "procedural" approach for a number of reasons. First, there is a connection between it and the well-known approach to normative matters taken by H. A. Simon, who used the same term in making a somewhat different distinction.[2] I have borrowed his term, and some of the insights of his approach, but have adapted both to my own purposes.

Secondly, the term "procedure" connotes a "temporally-sequenced process", and through using it I intend to suggest that man's rationality resides mainly in the process of his adaptation over time to the natural and social environment with which he interacts.

Finally, and related to the second point, the mechanical nature of a process or procedure is meant to suggest that man's rationality resides in properties that also arise in

our attempts as scientists to explain observed behavior. In both cases, normative and positive, it is through appreciating the dynamics of an agent's adaptation to a changing environment, that we are led to an understanding of his behavior and the cognitive life lying behind it.

The adoption of this approach raises a number of problems not usually confronted in normative theories, and a few of these are discussed in later sections. But there are two general sorts of questions regarding this approach which are best discussed now. First, since the procedural approach does not adopt the more traditional position that man's rationality is essentially a function of his ability to consciously compute, criticize and discuss, it must at some point show how these conscious, linguistic, problem-solving activities relate to the more fundamental "procedural" mechanism, and must suggest reasons why a cognitive terminology should not be adopted from the outset.

Secondly, since the procedural notion is <u>not</u> founded on these more or less uniquely-human abilities, man's rationality would seem to reside in properties shared with other sorts of creatures. Thus, adopting the procedural approach would seem to give rise to the conclusion that these "higher level" abilities do not make man any more "rational" than at least some more poorly equipped inhabitants of the planet.

These are both intriguing issues, and it is clear that they must eventually be faced by any proponent of the

procedural approach. However, we should not be misled into thinking that they are "problems" for that approach any more than for the major alternative, the dual-interpretation approach. As soon as it is agreed that man is rational, and that we are to understand his behavior in that light, then much the same issues will arise, though in a modified form.

To see this, consider the positive use of decision theory and game theory. A niggling background concern that has long bothered some economists is the fact that in real-life much human behavior is clearly not consciously deliberated.[3] When that is the case, the assumptions lying behind the explanatory employment of decision-theoretic or game-theoretic models would seem to be inappropriate. Too much sophisticated calculation is needed to generate such a model, let alone solve it, for it to be accepted in the explanatory domain without some sort of justification.

However economists respond to this challenge, it is clear that this problem and that raised by the first criticism of the procedural approach, are two sides of the same coin. At root is the question of man's rationality when his "higher level" abilities are not involved, and the procedural approach takes this statement of the problem seriously, concluding that with only a few limiting exceptions man always behaves rationally, simply because the true well springs of his rationality are more general abilities underlying his conscious activities. The development of

those higher level abilities, of course, gives rise to interesting normative and positive questions concerning the nature of learning, socialization, and the development of philosophical theories. But none of these problems are unique to the procedural approach.

As to the second implied criticism concerning the dubious rationality of animals, I would rather the procedural approach were viewed as attempting to answer important questions that have been ignored, not solved, by the dual-interpretation approach. The argument is simply that once some general theory of behavior (for example, a modified version of the theory of operant conditioning) is shown to lead to an improved understanding of human behavior and to an increase in our ability to modify and control it, then some degree of kinship between human and animal behavior has already been established. In this way the issue of animal rationality is raised for any approach which takes man to be rational, not just for the procedural approach.

These brief arguments, of course, do not respond directly to the puzzles raised by the two "criticisms" outlined above. They do, however, put them in some sort of perspective. It would seem that the concern lying behind them is the tension between the two models of man dominating our culture: the scientific-behavioral-causal model, and the humanistic-rational-cognitive model. The former adopts the viewpoint of the "observer" of behavior, and so, in general,

ignores man's conceptual abilities, while the latter adopts the viewpoint of the "agent", and so assumes without question that man's conscious, linguistic abilities are central and irreducible. The "scientist" maintains that the former is generally applicable; the "humanist" maintains that the latter applies uniquely to man.

The procedural approach tries to bridge the gap between these two schools of thought. This is part of the "unification" program outlined in Chapter I. By arguing that the rational model is, at its foundation, a scientific-observer model as well, it attempts a reconciliation between the two sides of what is probably the most fundamental split in our culture.

But the reconciliation should not be viewed as a "reduction", as so many "unifications" become. The objective is not to "eliminate" either viewpoint, but rather to demonstrate that they are basically one and the same model. There is no materialism lurking hidden in this identity theory.

Of course, such a unification is a long-range undertaking. In the present chapter I can only hope to point to the overriding concern, and show how one positive, theoretical framework of the scientific-observer sort has at its roots a model of behavior capable of being elaborated in both a cognitive and a behavioral manner. As remarked earlier, the actual development of those extensions in the positive and normative domains is left for future work.

9.3. Social Psychology and Exchange Theory

The theoretical framework of social exchange was developed during the past few decades to provide a conceptual structure capable of dealing with persistent problems in social psychology. The major contributors to this development have been George Homans,[4] Peter Blau,[5] John Thibault, Harold Kelly,[6] and Richard Emerson.[7]

The task these "theorists" confronted was the development of an "approach", "viewpoint", or "conceptual structure",[8] which could bridge the gap between psychological theories of behavior and small group theory, where the latter was viewed as the foundation for the theory of social structure. (Most saw this "bridge" resting on both sides of the gap, and so exchange theory is not essentially "reductionist" in spirit. George Homans, of course, represents the major exception to this rule.[9])

This chapter explores one relatively small, but important, contribution to this ongoing development -- the "structure" imposed on the exchange-theoretic framework by Richard Emerson. The major reason why Emerson's theory is chosen is that in two brief papers he has laid out the whole exchange-theoretic framework, from its operant-theoretic foundation, to the start of a theory of exchange networks, including status emergence, norm development, group formation, and so on.

In choosing Emerson, however, I do not want to give the

impression that his proposals have no history. This is clearly not the case. Emerson himself is quick to point out the intellectual debt owed to other theorists, most especially John Thibault and Harold Kelly.[10] To give complete credit, then, the structure I discuss below should be called the Thibault-Kelly-Emerson framework, but I will refer to it as, simply, Emerson's theory.

Chapter VII presented a brief outline of Emerson's approach to power-dependence relations in order to start preparing the ground for the introduction of his exchange theory. Although that discussion is not resumed here, the backwards linkage should be kept in mind. From Emerson's point of view it was the earlier work with social power that led him to exchange theory.[11] From the perspective of this chapter, the ability of Emerson's theory to provide a foundation for explanatory notions like social power is one of its most important features.

The focus of most of the studies in exchange theory, of course, is the dyad, since two-agent interactions provide the most natural bridge between psychology and small groups.[12] The same focus is adopted here, although it is recognized that the extension of the model to handle larger exchange networks is important future work.

One way of trying to examine social interaction in the dyad is to attempt to treat each agent's behavior in terms of purely psychological processes. This has, for the most part,

been rejected in social psychology for one simple reason. Psychological theories treat the environmental conditions surrounding individual behavior as givens, that is as independent variables. (Behavior is the dependent variable.) In studying the dyad, however, the behavior of one of the agents provides part of the environment determining the other's behavior, and vice-versa. Thus, neither behavior is an environmental, independent variable. Both are inter-dependent.[13]

To study behavior in a dyad, then, a theoretical structure is needed which recognizes and builds on this inter-dependence. Exchange theory develops such a structure on a foundation provided by operant theory.

Before proceeding to a discussion of this foundation, a few remarks are in order regarding the choice of operant theory, since some other, more "cognitive", psychological foundation might be thought to be preferable. First, as Emerson suggests, operant theory is more "...a method for controlling behavior than a theory about behavior".[14] In this way it is "atheoretical", and so is ideally suited to the sort of theoretical tampering exchange theory undertakes.

Secondly, the central feature of operant theory, reinforcement, can be looked upon as a form of "exchange with the environment". In the dyad, agents "exchange" behavior; in operant theory, we can view the "exchange" as being with a neutral environment and in return for a reinforcing

stimulus of some sort. In both types of exchange, moreover, the important effects are the "regulatory" ones. That is, the effects each exchange, or transaction, has on subsequent ones between the two "parties".

This way of viewing things places behavior in a temporally-extended setting never before considered in operant theory. With this shift in focus, the "exchange-relation" that develops between the agents becomes more important than each of the transactions occurring within the sequence. In operant terms, each agent becomes a "conditioned reinforcer" for the other. In cognitive terms, each agent comes to "value" the other for the rewards he mediates.

On the normative side, then, if a choice rule can be shown to be operative within Emerson's theory, it will be of a somewhat different form than the usual Bayesian rule.[15] The placing of behavior in a temporally-extended sequence does not remove the element of choice, but it does raise new sorts of problems.

On a more particular level, it is important to note that the notion of an "operant", or "unit of behavior", can be operationalized at any convenient level of complexity from a lever press to a corporation merger. This flexibility is built into operant theory by its generality, and it is essential to the generation of the theory of social exchange.[16]

At the same level of detail, it will be convenient in what follows to consider reinforcers as being partitioned into

what might be called "equivalence classes", or "domains". The boundaries between these domains of equivalent reinforcers are to be determined by the following rule:

> Stimulus y_k is of the same domain as stimulus y_m, if reinforcement by y_k increases satiation for y_m and vice versa.[17]

Now it is clear that this simple rule does not always lead to either a true "paritioning" of the class of all stimuli, nor does it necessarily lead to groups of pre-systematically "equivalent" stimuli. One stimulus may increase satiation in more than one "domain", and two members of a given "domain" may produce satiation to different degrees. Nevertheless, for the sake of convenience this simplification will be adopted.

Finally, throughout the discussion of the operant-theoretic foundation of Emerson's framework, the "situation" of the agent, that which either is or contains a "discriminative stimulus", will be treated perfectly generally. In focusing on the dyad, of course, social exchange theory deals with those cases where the "situation" is another agent, and the behavior they "exchange" is mutually reinforcing.

While this "social" interpretation of the foundation Emerson constructs is clearly of ultimate concern, the more general model is emphasized in much of what follows. Where appropriate, of course, examples are given to illustrate how the principles he develops apply to the dyad.

9.4. The Operant Foundation of Exchange Theory

Emerson's theory of social exchange is "constructed" on an operant-theoretic foundation, where, as noted earlier, the notion of "construction" means that the psychological principles have provided important "building blocks" for the concepts and principles of the sociological theory.[18] The basic elements of the construction of this foundation will now be discussed. Further elaboration of this foundation, and the presentation of a normative interpretation will be undertaken in the next section.

There are four primitive psychological terms underlying Emerson's theory: Actor (symbolized A, B, etc.), Behavior (symbolized x_1, x_2, etc.), Situation (symbolized S_1, S_2, etc.), and Stimulus (symbolized y_1, y_2, etc.). These primitive operant concepts combine to define the "smallest independently meaningful unit" for social exchange theory -- the notion of an exchange relation. Emerson defines this in the following way.

Let "prob(y_k)" be the probability that stimulus y_k is emitted in situation S_j when actor A performs action x_i. Let "prob(x_i)" be the probability that A performs x_i given that he is in situation S_j, and that y_k happens to be a "stimulus consequence" of x_i. (That is, it is more likely to be emitted in S_j when A performs x_i than when he does not.) Now, if we assume that y_k is a positive reinforcer for A, and that in S_j, y_k is a stimulus consequence of x_i, then:

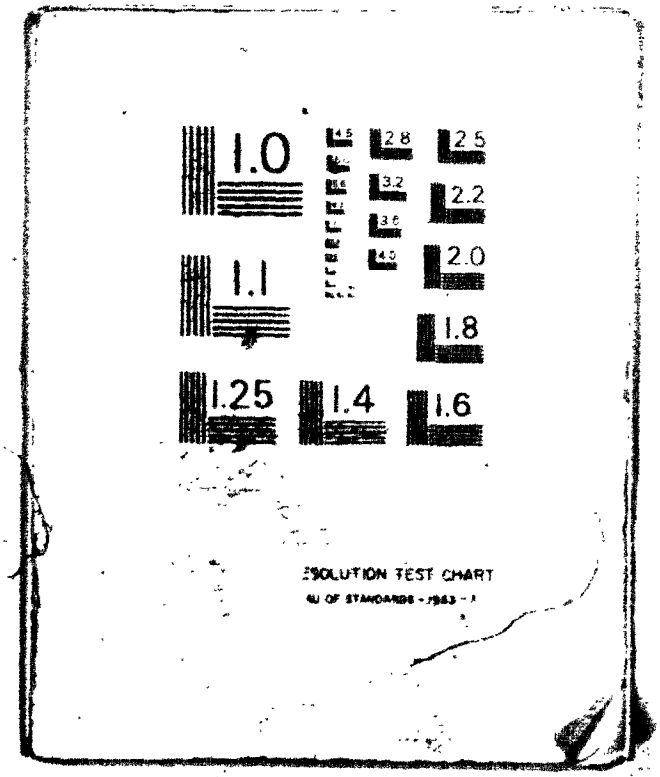
> ... prob(y_k) and prob(x_i) jointly define an organism-environment exchange relation (symbolized $Ax_i;S_jy_k$, or simply $A;S_j$ with x and y understood) consisting of a series of temporally interspersed opportunities (S_{jt}), initiations ($S_{jt} \rightarrow Ax_{it}$), and transactions ($(S_{jt} \rightarrow Ax_{it}) \rightarrow y_{kt}$); where $\rightarrow$ means "produces", evokes", or "is accompanied by".[19]

This definition is the "principle coordinating definition, bridging operant psychology with...social exchange theory".[20] As noted earlier, the bridge head is extended by introducing a second actor (B) in place of the discriminative stimulus S_j, and then considering transactions (i.e. behavior exchanges) which are reciprocally reinforcing. Thus the social exchange relation is between actors A and B (written A;B), and the actors' behavior repetoires contain items which can be "exchanged", and so, over time, lead to each actor being dependent on the other for valued rewards.

The development of Emerson's theory will now be presented. Where appropriate the possibility of a normative interpretation of his theory is discussed. Following this, the structure his concepts give to the normative theory is discussed, and I outline the work that lies ahead.

4 4

OF/DE


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

then the prob(x_i) are ordered as in Proposition 1, but in this case the ordering corresponds to a dependence ranking across exchange relations, not a "probability of reinforcement" ranking within an exchange relation.

On the positive side Emerson backs up this principle of choice behavior by treating dependence as a measure of the "situation" S_j's strength as a conditioned reinforcer.[28] In doing this, of course, he introduces considerations of social structure ignored by operant theory. For example, he considers the strength of the conditioned reinforcer, S_j, to be a function of the "alternative sources" the agent has for the rewards mediated by S_j, and no operant theory considers anything like this parameter.

But more to the point, in pursuing this connection with the operant foundation, Emerson makes dependence a function of three other parameters more closely linked to cognitive terminology. These are: the "value" of a reinforcer, the "comparison level" of an exchange relation, and the exchange relation's "primacy" (the number of reward domains it mediates).[29]

Although Emerson "...makes no attempt to weight the variables contributing to dependence, for that is an empirical matter",[30] it is clear that for normative purposes this is an unhappy resting point. As has been suggested in a number of places above, even if it is accepted that the procedural approach applied to Emerson's framework will lead to interesting

choice as that offered by Proposition 1 and his notion of "value", but the structure they offer is bare. Before an interesting normative theory can be extracted, then, much more empirical work is necessary to flesh it out. Only then can the theory of social exchange be seen to be the new and exciting focus for the whole of our cognitive vocabulary which I believe it to be.

9.6. Future Work

It would be redundant to end this chapter with a list of the open problems confronting proponents of the procedural approach. Much of the discussion has centered on exactly these points. Nevertheless, there are three specific suggestions I want to make, for each ties in with earlier sections of the dissertation.

The first suggestion concerns the use of exchange theory as an alternative to the normal form of representation. This has clearly been in the background all along, and it is hoped this connection can be pursued.

What this points to is the extension of the "hill-climbing" model of interaction discussed in Chapter IV along exchange-theoretic lines. During successive interactions, agents explore the space of possible interactions between them, and might even be thought of as optimizing some sort of "potential field" defined over that space.

The second suggestion concerns the development of

The empirical interpretation of Proposition 1 describes a relationship which holds between a set of objective reinforcement frequencies and certain behavioral frequencies, or, in the single-shot case, a certain "mixed strategy". Now, for the normative interpretation a Bayesian would demand (and Emerson thinks this is feasible),[23] that we interpret these $\text{prob}(y_k)_i$ as the agent's subjective probability estimates at a particular stage of the learning process. If this is done, the normative interpretation of Proposition 1 faces a dilemma. If the only relevant part of the agent's utility function is his utility for y_k, then by definition the agent is not acting rationally unless he chooses the optimal act, x_n.[24] On the other hand, importing other sorts of utility considerations would seem to be beside the point, since the proposition makes no mention of them. In either case, the straightforward extraction of cognitive notions (i.e. subjective probabilities and utilities) would seem to be in trouble.

It is argued below that, in general, Emerson's theory is not fully enough developed for the easy extraction of counterparts to the elements of the cognitive model, and so it is certainly true that caution is advised with Proposition 1. Nevertheless, in this case there are other sorts of considerations which enter the picture, and they are <u>not</u> "beside the point", whatever the Bayesian may think. In fact, what they point to is that through treating "learning" as just another element in a utility function, the Bayesian has

misconstrued the nature of the action itself.

To see this, consider, first, that as the agent enacts some act x_i, he not only obtains (or does not obtain) the reward y_k, he also adjusts the probability estimate $\text{prob}(y_k)_i$. (In operant terms x_i is positively reinforced.) But since the agent is also concerned with learning about his environment, the effect each "exchange" has on his probability estimates is an important factor which reflects itself in his choice behavior. That is, since in the situation defined by Proposition 1 the only way the agent can gather data to adjust his probability estimates is to enact a variety of non-optimal acts at various times, there would appear to be some plausibility to the normative model admitting some sort of mixed strategy during the learning stage.

Lest the Bayesian cry, "Foul!", at this point, and argue that the importance of "learning" is not included in the specification of the problem, let me point out that such learning is not only part of the agent's "values", it is also included in the notion of action! This is the import of Emerson's introducing the exchange relation as the "smallest independently meaningful unit" of his theory.[25] The "feedback" from the environment onto the agent's belief function after his performing actions is conceptually linked to the performance itself when we construe actions as taking place in the context of temporally-extended exchange relations. The notion of an exchange relation was defined as it was in order

to deal with this feedback process of adaptation. Consequently the importance of "learning" is not simply "another utility", according to the normative interpretation of Proposition 1, it is part of the way the rational agent thinks about the actions he is about to perform.

But Proposition 1 only considers the choice of such action when the single variable is the probability of reinforcement by a given stimulus. There are other general sorts of "choices", of course, including choice of the timing of initiations, choice among alternative exchange relations when the reinforcing stimulus varies, and choice among exchange relations simpliciter. Emerson's framework of principles deals with these on the foundation offered by Proposition 1, and the definition of an exchange relation.

Consider first, and briefly, the "timing" of initiations. Emerson follows traditional operant theory in suggesting that the timing of initiations (i.e. the frequency of initiations given continuous opportunities) is a direct function of the level of deprivation-satiation. That is, it is a function of the "...number and magnitude of transactions [in the same domain] during 'some' time period immediately preceding [time] t".[26] But, however interesting this phenomenon is on the positive side, there would seem to be little point to pursuing it on the normative side at the present time. Whether the effects of moment-to-moment deprivation-satiation are to be reflected in "values", or in some other cognitive dimension

is not at all clear. That true "deprivation" is reflected in "needs", not "wants", might be a clue here, but it is too early to tell. In any event, there are more important "choice points" to be considered, and so "timing" will have to be put aside.

The remaining choice points can be treated in two ways. The most general approach might be to consider the choice of behavior by an agent when simultaneous opportunities are presented in a variety of exchange relations each having a variety of potential transactions, and then develop a normative choice rule to characterize that behavior. The less general way would be to consider choice of exchange relation when simultaneous opportunities are presented, and let behavioral-choice within an exchange relation be governed by deprivation-satiation, Proposition 1, the initiations of the "other agent" and the balancing operations of Chapter VII. In what follows Emerson's approach is adopted, and choice among exchange relations is the focus for the general choice rule.

The empirical principle Emerson proposes is that when an agent, A, is faced with choice among a set of exchange relations $A;S_1, A;S_2, \ldots, A;S_n$, his initiation behavior will be such that the probability of his initiating a transaction in a given relation, S_j, varies directly with his dependence on S_j.[27] In other words, if we consider each exchange relations to have "formed" around a specific behavior $(x_1, x_2 \ldots, x_n)$

then the prob(x_i) are ordered as in Proposition 1, but in this case the ordering corresponds to a dependence ranking across exchange relations, not a "probability of reinforcement" ranking within an exchange relation.

On the positive side Emerson backs up this principle of choice behavior by treating dependence as a measure of the "situation" S_j's strength as a conditioned reinforcer.[28] In doing this, of course, he introduces considerations of social structure ignored by operant theory. For example, he considers the strength of the conditioned reinforcer, S_j, to be a function of the "alternative sources" the agent has for the rewards mediated by S_j, and no operant theory considers anything like this parameter.

But more to the point, in pursuing this connection with the operant foundation, Emerson makes dependence a function of three other parameters more closely linked to cognitive terminology. These are: the "value" of a reinforcer, the "comparison level" of an exchange relation, and the exchange relation's "primacy" (the number of reward domains it mediates).[29]

Although Emerson "...makes no attempt to weight the variables contributing to dependence, for that is an empirical matter",[30] it is clear that for normative purposes this is an unhappy resting point. As has been suggested in a number of places above, even if it is accepted that the procedural approach applied to Emerson's framework will lead to interesting

normative insight, more must be forthcoming before a cognitive choice rule can be extracted from the principles he proposes.

Be that as it may, the bare structure imposed by these four concepts does provide some guidelines, and I want to examine it a bit more closely before closing. In particular, Emerson's notions of "value" and comparison level" should be examined, for they point to a central problem with extracting the cognitive model.

The problem of attaching a "value" to a reinforcer has long bothered exchange theory. The obvious option is simply to let value be operationally equivalent to level of deprivation-satiation and be done with it. This option was taken by George Homans.[31] Emerson did not take this route, however, and for good reasons.

First, it is a common empirical observation that even when level of deprivation-satiation is held constant, reinforcers have varying abilities to evoke initiations. This in itself supports the conclusion that "something else" lies behind choice behavior than deprivation-satiation. Of course, this should not be surprising, since in most cognitive theories the notion of value is clearly not reducible to current state of deprivation, anyway. Even revealed preference theory in its most behaviorist form could not hope to make sense of behavior in terms of preference if all that lay behind choice behavior was moment-to-moment states of

deprivation.

For empirical and methodological reasons, then, some other psychological foundation for the notion of "value" would seem to be needed.

Emerson responds to this problem by defining the relative "value" of a domain of equivalent reinforcers (or the relative "value" of some magnitude level in a given domain) as "...the strength of that [domain or magnitude] in evoking and reinforcing initiations relative to other comparable [domains or magnitudes], holding deprivation constant and greater than zero".[32]

But no mere definition of this sort says what the psychological foundation is for the phenomenon of "value", and so Emerson proceeds with the following. The value of a reinforcer (y) is a direct function of the "uncertainty" of reinforcement of the "corresponding class of initiations". That is, the "value" of y is related to the problematic reinforcement of the behavior undertaken with the expectation of reward by y, over some recency-weighted history of exchange.[33]

In short, Emerson's suggestion is that what operates in the determination of the strength of reinforcers over and above the level of deprivation-satiation, is a history of problematic reinforcement, where how problematic the reinforcement has been is some non-linear function of prob(y) peaking somewhere between 0 and 1.[34] If in the past an agent has

found water to be plentiful but food scarce, then, even with deprivation held constant, he will be more likely to initiate transactions aimed at obtaining food than alternative transactions aimed at obtaining water. But this is not to say that water and food could not be equally rewarding to the agent once each is obtained. Which commodity is more rewarding is a matter Emerson wants to distinguish from that of "value", and this leads to some confusion.

As Emerson defines it, his notion of "value" is related directly to choice behavior, that is, to "initiation probability". Since the cognitive model has choice behavior arising from some combination of an agent's estimates of the likelihood of his receiving a reward and the importance of the reward to him, Emerson's notion must be viewed as already combining these factors according to some unknown rule. But, if so, how we might extract those factors and his principle of combination is not clear.

Emerson's concern is with "value in action" not "symbolic projections",[35] and that is well and good. But ours must be with "beliefs" and "cognitive values", too. A greater problematic history of reinforcement might lead to the development of a greater "cognitive value", but it would surely also lead to a relatively lower estimate of the probability of success of the corresponding action, and the influence of the latter on that action would be the reverse of the one predicted by Emerson.

There is clearly something odd about Emerson's treatment of "value" if it is thought of in cognitive terms. This might be why he backed away from the problem of translating it into that vocabulary. Once more, we are forced to follow his lead, but this time it is with great reservation. This tension must be cleared up before the extraction of the normative model can proceed any further. However, as that problem is partly empirical, it must be left to the scientists for now.

The other fairly obvious "cognitive" notion Emerson introduces is the "comparison level" of exchange relations.[36] If two exchange relations are "alternatives" to one another, then their comparison levels are the magnitudes of the reinforcers they offer. (In turn, these are, for quantitative reinforcers, the actual magnitudes over some recency-weighted period, or, for qualitative reinforcers, the probability of reinforcement, similarly weighted.) If the relations are not alternatives, however, their appropriate "comparison levels" are the "values" of the (different) reinforcers they offer. "Primacy" is then introduced to reflect the number of domains offered by each relation, and choice proceeds according to the general dependence rule discussed earlier.

Once again, however, Emerson fails to give any detail on the connections among these terms. As his proposals regarding "comparison levels" stand now, they offer the same promise of an interesting new focus for a normative theory of

choice as that offered by Proposition 1 and his notion of "value", but the structure they offer is bare. Before an interesting normative theory can be extracted, then, much more empirical work is necessary to flesh it out. Only then can the theory of social exchange be seen to be the new and exciting focus for the whole of our cognitive vocabulary which I believe it to be.

9.6. Future Work

It would be redundant to end this chapter with a list of the open problems confronting proponents of the procedural approach. Much of the discussion has centered on exactly these points. Nevertheless, there are three specific suggestions I want to make, for each ties in with earlier sections of the dissertation.

The first suggestion concerns the use of exchange theory as an alternative to the normal form of representation. This has clearly been in the background all along, and it is hoped this connection can be pursued.

What this points to is the extension of the "hill-climbing" model of interaction discussed in Chapter IV along exchange-theoretic lines. During successive interactions, agents explore the space of possible interactions between them, and might even be thought of as optimizing some sort of "potential field" defined over that space.

The second suggestion concerns the development of

problem-solving algorithms and their relationship to the rationality notion defined by Proposition 1. What this puzzle opens up to investigation is the connection between algorithms per se, and the use of them by rational agents. It might turn out, for example, that algorithm-using agents behave rationally to the extent that they value using algorithms, and we would be wrong to look upon their behavior as rational relative to the "symbolic" values they think they are employing in the calculation.

However that may turn out, it certainly raises in a most direct fashion the problematic connection between our concepts and their correlates -- the problem of representation. This problem is the root of the tension between "values in action" and "symbolic projections" which worried Emerson, and it is the root of the tension between the "scientific" model of man and the "humanistic" model of man alluded to earlier. If ever there was a nexus of philosophical problems worth working on, this is it. Unfortunately, it, too, must be left for future work.

FOOTNOTES

1. In particular his (1972a).

2. See Simon's (1957), and the March-Simon collaboration, Organizations (1955), especially the discussion of the distinction between substantive planning and procedural planning on pp. 140-141.

3. In response to this, some economists adopt the position that the "realism" of the assumptions lying behind their models is not important. Others suggest that deviations from the assumptions are "randomly distributed", and so cancel out at the macro level. It has also been suggested that the rational model is generally applicable since most economic behavior in the marketplace is consciously computed, and that that is so because most economic agents are in a position to have to justify their actions relative to commonly recognized goals (eg. profit maximization).

4. See his (1974).

5. See his (1964).

6. See Thibault's and Kelly's (1959).

7. See his (1962), (1969), (1972a) and (1972b).

8. Both Thibault and Kelly and Richard Emerson emphasize this "conceptual" nature of their studies. See Thibault and Kelly (1959), p. vii, and Emerson's (1972a), pp. 38-39.

9. See Emerson's (1972a), p. 39. His reductionism is well known.

10. Emerson (1972a), p. 38.

11. Ibid., p. 39.

12. For a more detailed explanation of this tendency to focus on the dyad, see Thibault and Kelly (1959), p. 6.

13. Ibid., p. 2.

14. Emerson (1972a), p. 42.

15. In what follows I refer to the "Bayesian" rule in order to give a label to that more conventional interpretation of rational choice in terms of subjective probabilities and the maximization of expected utility.

16. See Emerson (1972a), p. 49-50.

17. This is a summary of the discussion in Ibid., p. 50.

18. See Emerson (1969), p. 403 for a discussion of this notion.

19. Emerson (1972a), p. 45.

20. Ibid., p. 45.

21. Ibid., p. 46.

22. Ibid., p. 47.

23. Ibid., p. 44.

24. This follows from the definition of utility.

25. As he puts it, operant, positive reinforcer, and discriminative stimulus constitute a "single conceptual unit", (Ibid., p. 44), and we should not break into it in an attempt to "explain" behavior.

26. Ibid., p. 47-48.

27. Ibid., p. 51.

28. Ibid., p. 48-50.

29. Ibid., p. 51-56.

30. Ibid., p. 57.

31. See Emerson's discussion in Ibid., p. 51.

32. Ibid., p. 53.

33. Ibid., p. 53-54.

34. Emerson "suggests" a functional relation, but there is no need to reproduce it here. See Ibid., p. 53-54 for his measure of "uncertainty".

35. Ibid., p. 55.

36. He follows Thibault and Kelly here. See their discussions in (1959), p. 21-23.

Rapoport, A. (ed.)
1974 Game Theory as a Theory of Conflict Resolution. Boston: Reidel.

Rapoport, A., and Chammah, A. M.
1965 Prisoner's Dilemma. Ann Arbor: University of Michigan Press.

Riker, W. H.
1964 "Some Ambiguities in the Notion of Power." The American Political Science Review LVIII: 341-349.

Schelling, Thomas.
1960 The Strategy of Conflict. Cambridge: Harvard University Press.

Schopler, J.
1965 "Social Power." In Advances in Experimental Social Psychology, Volume 2, edited by L. Berkowitz, pp. 177-218. New York: Academic Press.

Simon, H. A.
1957 Models of Man: Social and Rational. John Wiley and Sons.

Simpson, R. L.
1972 Theories of Social Exchange. New York: General Learning Press.

Stalnaker, Robert.
1968 "A Theory of Conditionals." In Studies in Logical Theory, edited by N. Rescher. Blackwell.

Thibaut, J. W., and Kelly, H. H.
1959 The Social Psychology of Groups. New York: John Wiley and Sons.

Vickers, G.
1972 Freedom in a Rocking Boat. Pelican Books.

von-Neumann, J., and Morgenstern, O.
1944 Theory of Games and Economic Behavior. John Wiley and Sons.

1947 Theory of Games and Economic Behavior. Second Edition. John Wiley and Sons.

Weber, Max.
1947 The Theory of Social and Economic Organization. Oxford University Press.

Emerson, R. M.
1969 "Operant Psychology and Exchange Theory." In Behavioral Sociology, edited by Burgess and Bushell, pp. 379-405. New York: Columbia University Press.

1972a "Exchange Theory, Part I: A Psychological Basis for Exchange Theory." Sociological Theories in Progress, Volume 2, edited by Berger et al., pp. 38-57. Boston: Houghton Mifflin Company.

1972b "Exchange Theory, Part II: Exchange Relations and Network Structures." Sociological Theories in Progress, Volume 2, edited by J. Berger et al., pp. 58-87. Boston: Houghton Mifflin Company.

French, J. R. P., Jr., and Raven B.
1968 "The Bases of Social Power." In Group Dynamics Research and Theory (3rd edition), edited by Dorwin Cartwright and Alvin Zander, pp. 259-269. Harper and Row.

Gauthier, David.
1975 "Reason and Maximization." Canadian Journal of Philosophy 4(3): 411-433.

Gergen, K. J.
1969 The Psychology of Behavior Exchange. Reading, Mass.: Addison-Wesley.

Gibbard, A., and Harper, W. L.
1976 "Counterfactuals and Two Kinds of Expected Utility." In Foundations and Applications of Decision Theory, edited by C. A. Hooker, et al. Reidel, forthcoming.

Goldman, A.
1970 A Theory of Human Action. Prentice-Hall.

1972 "Toward a Theory of Social Power." Philosophical Studies 23: 221-268.

1974 "On the Measurement of Power." Journal of Philosophy 71: 231-252.

Hardy, Rollo.
1964 Methodology of the Behavioral Sciences (Problems and Controversies. Springfield, Illinois: C. C. Thomas.

Harsanyi, J. C.

1956 "Approaches to the Bargaining Problem before and after the Theory of Games: A critical discussion of Zeuthan's, Hick's and Nash's theories." Econometrica, 24: 144-157.

1959 "A Bargaining Model for the Co-operative n-Person Game." In Contributions to the Theory of Games 4, edited by W. Tucker and R. D. Luce. Princeton: Princeton University Press.

1961 "On the Rationality Postulates Underlying the Theory of Co-operative Games." Journal of Conflict Resolution 5(2): 179-196.

1962a "Measurement of Social Power, Opportunity Costs, and the Theory of Two-Person Bargaining Games." Behavioral Science 7: 67-80.

1962b "Measurement of Social Power in n-Person Reciprocal Power Situations." Behavioral Science 7: 81-91.

1963 "A Simplified Bargaining Model for the n-Person Cooperative Game." International Economic Review 4: 194-220.

1964 "A General Solution for Finite Non-Co-operative Games, Based on Risk Dominance." In Advances in Game Theory, edited by M. Dresher et al., pp. 651-679. Princeton: Princeton University Press.

1966a "A Bargaining Model for Social Status in Informal Group and Formal Organizations." Behavioral Science 12: 357-369.

1966b "A General Theory of Rational Behavior in Game Situations." Econometrica 34: 613-634.

1968 "Individualistic and Functionalistic Explanations in the Light of Game Theory: The example of social status." In Problems in the Philosophy of Science, edited by I. Lakatos and A. Musgrave, pp. 305-332. Amsterdam, North-Holland.

1975a "The Tracing Procedure: A Bayesian Approach to Defining a Solution for n-Person Non-cooperative Games." Distributed by the author at the Fifth International Congress of Logic, Methodology and Philosophy of Science, August 1975, The University of Western Ontario, London, Canada.

Harsanyi, J. C.
1975b "The Tracing Procedure, Part II." Distributed by the author at the Fifth International Congress of Logic, Methodology and Philosophy of Science, August 1975, The University of Western Ontario, London, Canada.

1975c "The Tracing Procedure: A Bayesian Approach to Defining a Solution for n-Person Non-cooperative Games." International Journal of Game Theory 5.

Hollis, M., and Nell, E.
1975 Rational Economic Man. Cambridge University Press.

Homans, George C.
1974 Social Behavior: Its Elementary Forms. Revised edition. New York: Harcourt Brace Javonovich.

Howard, Nigel.
1971 The Paradoxes of Rationality. Cambridge: MIT Press.

Kuhn, A.
1963 The Study of Society: A Unified Approach. Irwin.

Leach, J.
1975 "The Dual Function of Rationality." Paper presented at the Fifth International Congress of Logic, Methodology and Philosophy of Science, August 1975, The University of Western Ontario, London, Canada. (To appear in the published proceedings of the Congress.)

Levi, I.
1975 "Newcomb's Many Problems." Theory and Decision 6: 161-175.

Lowe, A.
1970 On Economic Knowledge. Harper and Row.

Luce, R. Duncan.
"Conjoint Measurement." In Foundations and Applications of Decision Theory, edited by C. A. Hooker et al. Reidel, forthcoming.

Luce, R. Duncan, and Raiffa, H.
1957 Games and Decisions. New York: John Wiley and Sons.

Lukes, Stephen.
1973 "Methodological Individualism Reconsidered." In The Philosophy of Social Explanation, edited by Alan Ryan, pp. 119-129. Oxford University Press.

March, J. G.
1955 "An Introduction to the Theory of Measurement of Influence." The American Political Science Review 49: 431-451.

March, J. G., and Simon, H. A.
1958 Organizations. John Wiley and Sons.

McClennen, Edward, F.
1976a "Some Formal Problems with the von-Neumann and Morgenstern Theory of Two-Person Zero-Sum Games I: The Direct Proof." Theory and Decision 6.

1976b "The Minimax Theory and Expected Utility Reasoning." In Foundations and Applications of Decision Theory, edited by C. A. Hooker et al. Reidel, forthcoming.

McKinsey, J. C. C.
1952 Introduction to the Theory of Games. McGraw-Hill.

Meeker, B. F.
1971 "Decisions and Exchange." American Sociological Review 36: 485-495.

Narens, Louis.
1974 "Measurement without Archimedean Axioms." Philosophy of Science 41: 374-393.

Nash, J. F.
1950 "The Bargaining Problem." Econometrica 18: 155-162.

1953 "Two-Person Cooperative Games." Econometrica 21: 128-140.

Nozick, R.
1969 "Newcomb's Problem and Two Principles of Choice." In Essays in Honor of Carl G. Hempel, edited by N. Rescher, pp. 114-146. Dorchrecht: Reidel.

Rapoport, A.
1966 Two-Person Game Theory: The Essential Ideas. Ann Arbor: University of Michigan Press.

1970 N-Person Game Theory: Concepts and Applications. Ann Arbor: University of Michigan Press.

Rapoport, A. (ed.)
1974 Game Theory as a Theory of Conflict Resolution. Boston: Reidel.

Rapoport, A., and Chammah, A. M.
1965 Prisoner's Dilemma. Ann Arbor: University of Michigan Press.

Riker, W. H.
1964 "Some Ambiguities in the Notion of Power." The American Political Science Review LVIII: 341-349.

Schelling, Thomas.
1960 The Strategy of Conflict. Cambridge: Harvard University Press.

Schopler, J.
1965 "Social Power." In Advances in Experimental Social Psychology, Volume 2, edited by L. Berkowitz, pp. 177-218. New York: Academic Press.

Simon, H. A.
1957 Models of Man: Social and Rational. John Wiley and Sons.

Simpson, R. L.
1972 Theories of Social Exchange. New York: General Learning Press.

Stalnaker, Robert.
1968 "A Theory of Conditionals." In Studies in Logical Theory, edited by N. Rescher. Blackwell.

Thibaut, J. W., and Kelly, H. H.
1959 The Social Psychology of Groups. New York: John Wiley and Sons.

Vickers, G.
1972 Freedom in a Rocking Boat. Pelican Books.

von-Neumann, J., and Morgenstern, O.
1944 Theory of Games and Economic Behavior. John Wiley and Sons.

1947 Theory of Games and Economic Behavior. Second Edition. John Wiley and Sons.

Weber, Max.
1947 The Theory of Social and Economic Organization. Oxford University Press.

Williams, J. D.
1954 The Compleat Strategyst. New York: McGraw Hill.

Wilson, Bryan.
1970 Rationality. Harper and Row.

Zeuthen, F.
1930 Problems of Monopoly and Economic Warfare. London: George Routledge and Sons.